UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

E.A.P. DE ESTADÍSTICA

Medidas de influencia en el análisis discriminante lineal para dos grupos y algunas aplicaciones en la etnobotánica

TESIS

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Edilberto Daniel Cañari Casaño

ASESOR

Doris Gómez Ticerán

Lima - Perú

2010

A mis padres por
enseñarme a luchar
en la vida

Agradecimientos

El presente trabajo es el resultado de un enorme esfuerzo que tiene por finalidad cubrir uno de los vacios, en nuestro medio, como es el uso de medidas estadísticas para identificar observaciones influyentes en el marco del análisis discriminante lineal.

Quiero expresar mi agradecimiento a cada una de las personas que a continuación nombraré, quienes me ayudaron de manera desinteresada y muchas de ellas han sido un soporte muy fuerte para mí hasta el día de hoy.

A mis padres, porque gracias a su esfuerzo he podido terminar mis estudios y consiguientemente esta tesis, a pesar de las dificultades. A mis hermanos por su apoyo en especial a Miguel y María.

A mi asesora la Dra. Doris Gómez Ticerán, por motivarme y guiarme en este proyecto, por sus sabios consejos, por sus enseñanzas, estoy seguro que sin ellos, alcanzar esta meta hubiera sido más difícil.

Al profesor José Antonio Díaz de la Universidad de Granada-España, por su apoyo con algunos materiales bibliográficos y grandes consejos.

A los profesores de la Escuela Académico Profesional de Estadística, por ser los artífices de mi formación académico-profesional.

A mis amigos, por compartir grandes momentos en la vida universitaria y a todas las personas que me apoyaron para llevar a cabo este proyecto.

Al Consejo Superior de Investigaciones, del Vicerrectorado de Investigación de la Universidad Nacional Mayor de San Marcos, por el soporte financiero a este trabajo de tesis, a través del concurso: Fondo de Promoción de Tesis.

MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL PARA DOS GRUPOS Y ALGUNAS APLICACIONES EN LA ETNOBOTÁNICA

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para optar el Título Profesional de Licenciado en Estadística

Aprobada por:

_____

Mg. Olga Lidia Solano Dávila
Presidenta

_____

Lic. Grabiela Montes Quintana
Miembro

_____

Dra. Doris Gómez Ticerán
Miembro – Asesor

Lima – Perú

Enero 2010

FICHA CATALOGRÁFICA

CAÑARI CASAÑO, EDILBERTO DANIEL

Medidas de influencia en el análisis discriminante lineal para dos grupos y algunas aplicaciones en la etnobotánica, (Lima) 2010.

159p, 30 cm. (UNMSM, Licenciado en Estadística, 2010)
Tesina, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas. Estadística.
I. UNMSM / F. de C.M. II. Medidas de influencia en el análisis discriminante lineal para dos grupos y algunas aplicaciones en la etnobotánica.

RESUMEN

MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL PARA DOS GRUPOS Y ALGUNAS APLICACIONES EN LA ETNOBOTÁNICA

PRESENTADO POR: Bachiller. CAÑARI CASAÑO, Edilberto Daniel
DIRIGIDO POR: Dra. GÓMEZ TICERÁN, Doris

ENERO 2010

Se presentan diversas medidas para identificar observaciones influyentes en el marco del análisis discriminante lineal en dos grupos. Se simula una muestra de vectores aleatorios en el espacio R^4, en la que, con las medidas presentadas, se identifican observaciones influyentes. Se completa el trabajo, con algunas aplicaciones en la etnobotánica, en las que también se identifican observaciones influyentes. Previamente, se presenta los aspectos básicos del análisis discriminante lineal en dos grupos y se establece la relación de proporcionalidad entre los coeficientes de la función lineal discriminante y los coeficientes del ajuste del modelo de regresión lineal múltiple.

ABSTRACT

MEASURES OF INFLUENCE IN LINEAR DISCRIMINANT ANALYSIS FOR TWO GROUPS AND SOME APPLICATIONS IN THE ETHNOBOTANY

PRESENTADO POR: Bachiller. CAÑARI CASAÑO, Edilberto Daniel
DIRIGIDO POR: Dra. GÓMEZ TICERÁN, Doris

JANUARY 2010

It shows several measures for identifying influential observations in the framework of linear discriminant analysis into two groups. It simulates a sample of random vectors in space R^4, in which, with the measures presented, identify influential observations. Work is completed, with some applications in ethnobotany, which also identifies influential observations. Previously, it presents the basics of linear discriminant analysis into two groups and we establish the relationship of proportionality between the coefficients of the linear discriminant function and the coefficients for the adjustment of multiple linear regression model.

Keywords:
Influence measures, linear discriminant analysis for two groups, influential observations.

INTRODUCCIÓN

En todo conjunto de datos, generalmente existen algunas observaciones que tienen un comportamiento diferente a la gran mayoría de ellos, a los cuales diversos investigadores han llamado observaciones discordantes, contaminantes, outliers, solo por nombrar algunos términos que han sido asignados a tales observaciones, Beckman y R. Cook (1983) y que en muchas ocasiones, estas pueden afectar diversos aspectos del análisis estadístico. Para solucionar dicho problema, se han desarrollado de un gran número de estudios en los que se han propuesto un conjunto de métodos y/o medidas que sirven para detectar tales observaciones, que en general reciben el nombre de Análisis de Influencia, (Muños et al., 2001).

El Análisis de Influencia, en particular la identificación de observaciones influyentes, ha sido ampliamente estudiado y difundido en diversas aplicaciones del análisis de regresión, en esta línea argumental, cabe citar a los trabajos de Belsley et al., (1982)[1].

Uno de los tópicos más importantes para analizar observaciones multivariantes, se presenta en la literatura, con el título de análisis discriminante o discriminación y clasificación, cuyo interés principal es ubicar un individuo (objeto) o un grupo de ellos en una de las categorías, grupos o poblaciones concurrentes preestablecidas.

La clasificación consiste en la identificación del grupo al cual pertenece el individuo, llevando en cuenta sus características observadas. Cuando tales características son mediciones numéricas, la designación a los grupos se llama discriminación y la combinación de las mediciones recibe el nombre de función discriminante. Más específicamente, en la discriminación se pretende describir de manera gráfica (en dos o tres dimensiones) o algebraicamente mediante funciones llamadas funciones discriminantes, los aspectos que diferencian a los individuos u objetos de varias poblaciones, Anderson (1958).

[1] Referencia en Enguix (2001)

Los métodos más utilizados en problemas prácticos de discriminación y clasificación, son el de Fisher, el de razón de verosimilitud y el de Bayes, que producen funciones de discriminación lineales en el caso homocedástico, (Anderson,1958; Mardia,1970 y Lachenbruch, 1968), cuadráticas en el caso heteroscedástico, (Anderson,1958 y Mardia, 1970).

En principio, independiente de la condición de homoscedasticidad o heteroscedasticidad de las matrices de covarianzas de las poblaciones concurrentes, en los problemas de clasificación estamos sujetos a los denominados "errores de clasificación", esto es, clasificar un individuo en una población cuando en realidad procede de otra población, por lo que interesa evaluar la capacidad predictiva de la función discriminante y consecuentemente de la regla de clasificación. La identificación de observaciones influyentes, que como ya se señaló, ampliamente estudiado y difundido en diversas aplicaciones del análisis de regresión; en el marco del análisis discriminante no ha tenido mayor difusión, a pesar de haber sido abordado mediante diferentes enfoques. Campbell (1978) propuso medidas de influencia mediante la función de influencia propuesta por Hampel (1974); basándose en la aproximación de esta función, Cook y Weisberg (1982)[2] propusieron una medida de influencia para la probabilidad de mala clasificación; Fung (1995) apoyándose en la relación que existe entre los coeficientes de la función discriminante lineal de Fisher y los coeficientes del modelo de regresión lineal múltiple, propuso algunas medidas siguiendo la metodología usada en análisis de regresión. Sin embargo, el uso de estas medidas no han sido ni aplicadas y mucho menos difundidas en nuestro medio.

En el contexto descrito, el objetivo de este trabajo es presentar las diversas medidas desarrolladas, por los autores ya citados, para identificar observaciones influyentes en el marco del análisis discriminante lineal en dos grupos. Se simula una base datos con vectores aleatorios en el espacio R^4, en la que se pretende identificar las observaciones influyentes con las medidas mencionadas. Se completa el trabajo, con algunas aplicaciones en la etnobotánica, en las que también se identifican observaciones influyentes.

Para alcanzar el objetivo planteado, el presente trabajo se ha dividido en cuatro capítulos. En el primer capítulo se presenta aspectos básicos sobre dos métodos clásicos, el análisis discriminante lineal y el análisis de regresión lineal múltiple, también se presenta el desarrollo teórico y metodológico de la relación que existe entre los coeficientes de ambos métodos,

[2] Referencia Fung(1992).

finalmente se muestra tres aplicación sobre la relación existente entre los coeficientes de ambos métodos, mediante una aplicación en la etnobotánica. En el segundo capítulo se presenta aspectos importantes sobre las observaciones discordantes, obtenidos de diversos estudios realizados sobre tales observaciones, además se hace el estudio sobre el efecto de una observación discordante en las estimaciones de los parámetros de posición central y de dispersión mostrado en Peña (2000), porque permite cuantificar el efecto de una observación discordante en las estimaciones de los parámetros involucrados en el análisis discriminante lineal.

En el tercer capítulo, se presenta una revisión sobre los estudios realizados en el análisis de influencia, una descripción sobre la función de influencia propuesta por Hampel (1974), así como la aproximación de dicha función propuesta por Devlin et al (1975)[3], también se presenta el fundamento teórico y el desarrollo metodológico de medidas de influencia para diversos parámetros del análisis discriminante lineal para dos grupos. Finalmente, mediante datos simulados se valida el potencial de las medidas de influencia presentadas en la sección anterior.

En el cuarto capítulo, se presentan tres aplicaciones en la biología, dos de las cuales corresponden a la botánica. El primer conjunto de datos cumple con los supuestos de homoscedasticidad de las matrices de covarianzas de las poblaciones concurrentes, mientras que en las dos siguientes aplicaciones las matrices de covarianzas son diferentes; sin embargo, como se verificará, aún así, ha sido posible la identificación de observaciones influyentes, a pesar que los datos no cumplen con el supuesto de homoscedasticidad de las matrices de covarianzas.

[3] Referencia en Campbell(1978); Enguix(2001); Fung(1992).

ÍNDICE

CAPITULO III

MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL EN DOS GRUPOS

CAPITULO IV

APLICACION DE LAS MEDIDAS DE INFLUENCIA

CAPÍTULO I

ANÁLISIS DISCRIMINANTE LINEAL Y ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

1.1. INTRODUCCIÓN

En investigaciones aplicadas, es muy común encontrar situaciones, en las que se debe estimar o analizar el comportamiento de una variable dependiente, en función de una o más variables independientes. Cuando la variable dependiente es cuantitativa, se suele hablar de problemas de predicción o estimación, que bajo ciertas condiciones el análisis de regresión es uno de los métodos más aplicados en tales situaciones. Cuando la variable dependiente es cualitativa o categórica, y con las independientes se construye una función lineal a fin de ser usada para clasificar individuos, el análisis discriminante, bajo ciertos supuestos, es un método que ofrece buenos resultados. En ambos métodos y en forma independiente, los datos deben cumplir una serie de supuestos, entre las que destacan: normalidad e igualdad de varianzas, independencia, linealidad entre otros. A pesar de las dos métodos son filosóficamente diferentes, (Anderson, 1984) probó que existe una relación entre los coeficientes de ambos métodos, cuyo desarrollo metodológico se presenta en detalle, con el fin de justifica en uso de medidas de influencia en el análisis discriminante lineal, ampliamente usado en regresión lineal múltiple.

Con el objetivo de uniformizar la notación a lo largo del trabajo, en el presente capítulo se presentará aspectos básicos sobre el análisis de regresión lineal múltiple y el análisis discriminante lineal. En la primera sección se presenta la teoría básica sobre el análisis discriminante lineal, las distintas reglas de clasificación y la manera de estimar las probabilidades de mala clasificación. En la segunda sección se presenta los aspectos básicos sobre el análisis de regresión lineal múltiple. En la tercera sección se presentará el desarrollo teórico y metodológico, sobre la relación que existe entre los coeficientes de ambos métodos. Finalmente se muestra una aplicación para ilustrar los resultados teóricos.

1.2. EL PROBLEMA GENERAL DE DISCRIMINACIÓN

1.2.1. Introducción

Sean G_k dos poblaciones o clases de sujetos y $\vec{X}^{(k)^T} = \left(X_1^{(k)},...., X_p^{(k)}\right)$, $k = 1,2$, un vector aleatorio con valores en R^p conteniendo mediciones de los individuos de cada una de las poblaciones. Los valores observados de $\vec{X}^{(k)}$ difieren de una clase a otra y a través de sus mediciones o medidas, $\vec{x}^{(k)^T} = \left(x_1^{(k)},...., x_p^{(k)}\right)$, construiremos una regla para clasificar un nuevo individuo $\vec{x}^T = \left(x_1, x_2,..., x_p\right)$ de R^p en una de las dos poblaciones.

En la construcción de una función discriminante, la suposición de igualdad de matrices de covarianzas de las poblaciones concurrentes (homoscedasticidad) es una suposición fuerte y que en muchas aplicaciones no es satisfecha.

Bajo el supuesto de homoscedasticidad se obtiene una función lineal (Lachenbruch,1975), mientras que sobre el supuesto de heteroscedasticidad la función discriminante contiene el término cuadrático (Anderson, 1984).

En principio, independientemente de la suposición de homoscedasticidad o heteroscedasticidad de las matrices de covarianzas de las poblaciones concurrentes, en el problema de clasificación estamos sujetos a los posibles errores de clasificación, esto es, ubicar a un individuo en cierta población cuando en realidad pertenece a otra. Por tanto, en la construcción de procedimientos de clasificación se busca aquella que minimiza el costo esperado de mala clasificación (Anderson,1984).

1.2.2. Reglas de buena clasificación en dos poblaciones

Podemos pensar en una observación, como un punto en el espacio de dimensión p . Dividimos tal espacio en dos regiones disjuntas R_1 y R_2 $\left(R_1 \cup R_2 = R^p\right)$. Si la observación

$\vec{x}^{T} = (x_1, x_2, ..., x_p)$ pertenece a R_1 es clasificado como procedente de G_1 y si pertenece a R_2 es clasificado como perteneciente a G_2, o sea:

Clasificar $\vec{x}$ en G_k si $\vec{x} \in R_k \quad k = 1,2$ (1)

con $R_1 \cup R_2 = R^p$ y $R_1 \cap R_2 = \phi$

Las poblaciones $G_k, k = 1,2$, están caracterizadas por sus funciones de densidad $f_1(\vec{x})$ y $f_2(\vec{x})$ respectivamente.

En todo proceso de clasificación se tiene asociado el costo, dado que, cuando clasificamos un objeto en G_1 cuando en realidad pertenece a G_2, o lo clasificamos en G_2 dado que procede de G_1, se comete un error que tiene un costo asociado. Dichos costos pueden medirse en cualquier unidad, porque lo que importa es el cociente entre ellos. Claramente, un buen procedimiento de clasificación es aquel que minimiza el costo de mala clasificación.

Para la posterior presentación, asumiremos la siguiente notación:

$f_k(\vec{x})$ Función de densidad de probabilidad del vector de observaciones $\vec{x}$ correspondiente a la población G_k

R Regla de clasificación particular

R_k Región de clasificación correspondiente a la población G_k

π_k Probabilidad a priori, de obtener una observación de la población o grupo G_k;

C(2/1) costos de clasificar equivocadamente la observación $\vec{x}$ de G_1 en el grupo G_2.

C(1/2) costos de clasificar equivocadamente la observación $\vec{x}$ de G_2 en el grupo G_1.

$P(1/2;R)$ Probabilidad condicional de clasificar equivocadamente la observación $\vec{x}$ en el grupo G_1, cuando realmente procede del grupo G_2 según la regla R.

$P(2/1;R)$ Probabilidad condicional de clasificar equivocadamente la observación $\vec{x}$ en el grupo G_2, cuando realmente procede del grupo G_1 según la regla R.

$r(j,R)$ riesgo o pérdida esperada, cuando una observación de la población G_j es clasificada en la población G_i.

$$P(\vec{x} \text{ ser clasificado correctamente}) = \int_{R_k} f_k\left(\vec{x}\right) d\vec{x}.$$

$$P(\vec{x} \text{ ser clasificado incorrectamente}) = \int_{R_i} f_j\left(\vec{x}\right) d\vec{x}., \text{ donde, } \quad d\vec{x} = dx_1 dx_2 .... dx_p$$

$$P(2/1;R) = 1 - P(1/1;R) \qquad P(1/2;R) = 1 - P(2/2;R)$$

$$r(1,R) = C(2/1)P(2/1,R) \qquad r(2,R) = C(1/2)P(1/2,R)$$

Definición 1 El costo esperado de mala clasificación (CEM) de la regla R, (Anderson, 1984), se define como la suma de los productos de la probabilidad de que una observación pertenezca a una determinada población por la pérdida esperada de la misma. O sea:

$$CEM = \pi_1 r(1,R) + \pi_2 r(2,R) = \pi_1 C(2/1)P(2/1,R) + \pi_2 C(1/2)P(1/2,R) \qquad (2)$$

En procedimiento que minimiza (3) se llama procedimiento de Bayes para las probabilidades a priori π_1 y π_2. Todo procedimiento de Bayes es admisible, es decir, en la clase de procedimientos R no existe otro mejor que él (Anderson, 1984).

Definición 2 La probabilidad total de mala clasificación (PTM) de la regla R, (Anderson, 1984), se define como:

PTM = $P(\vec{x}$ ser clasificad o incorrecta mente en G_1 o $G_2)$

$$= \pi_1 P(2/1,R) + \pi_2 P(1/2,R)$$

$$= \pi_1 \int_{R_2} f_1\left(\vec{x}\right) d\vec{x} + \pi_2 \int_{R_1} f_2\left(\vec{x}\right) d\vec{x} \qquad (3)$$

Cuando se conocen las distribuciones de probabilidad, debe usarse tal información para la construcción de las reglas de clasificación.

Regresemos al problema de escoger las regiones R_1 y R_2 tal que el CEM (2) sea mínimo. Dada la regla R, de (2) tenemos:

$$\text{CEM} = \pi_1 \; C(2/1)P(2/1,R) + \pi_2 \; C(1/2)P(1/2,R)$$

$$= \pi_1 \; C(2/1) \int_{R_2} f_1\left(\vec{x}\right) d\vec{x} + \pi_2 \; C(1/2) \int_{R_1} f_2\left(\vec{x}\right) . d\vec{x}$$

$$= \pi_1 \; C(2/1) \left\{ 1 - \int_{R_1} f_1\left(\vec{x}\right) . d\vec{x} \right\} + \pi_2 \; C(1/2) \int_{R_1} f_2\left(\vec{x}\right) . d\vec{x}$$

$$= \pi_1 \; C(2/1) + \pi_2 \; C(1/2) \int_{R_1} f_2\left(\vec{x}\right) . d\vec{x} - \pi_1 \; C(2/1) \int_{R_1} f_1\left(\vec{x}\right) . d\vec{x}$$

$$\text{CEM} = \pi_1 \; C(2/1) + \int_{R_1} \left[\pi_2 C(1/2) f_2(\vec{x}) - \pi_1 C(2/1) f_1(\vec{x}) \right] d\vec{x} \qquad (4)$$

La expresión (4) es mínima cuando la región R_1 se escoge de manera que:

$\pi_2 C(1/2) f_2(\vec{x}) - \pi_1 C(2/1) f_1(\vec{x}) \le 0$,o equivalentemente:

$$\frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)} \ge \frac{\pi_2 C(1/2)}{\pi_1 C(2/1)} \qquad (5)$$

Así, la regla que minimiza el costo esperado de mala clasificación (CEM) se resume en el siguiente teorema.

Teorema 1. Sean π_1 y π_2 las probabilidades a priori de clasificar una observación de la población G_1 con densidad $f_1(\vec{x})$ y de la población G_2 con densidad $f_2(\vec{x})$, y si $C(2/1)$ es el costo de clasificar una observación de G_1 como de G_2 y $C(1/2)$ es el costo de clasificar una observación de G_2 como de G_1 ; entonces, las regiones de clasificación R_1 y R_2, definidas por:

$$R_1: \quad \frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)} \geq \frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}$$

$$R_2: \quad \frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)} < \frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}$$

minimizan el costo esperado de mala clasificación.

Mientras que la regla que minimiza la probabilidad total de mala clasificación (PTM), se resume en el siguiente teorema.

Teorema 2. Sean π_1 y π_2 las probabilidades a priori de retirar una observación de la población G_1 con densidad $f_1(\vec{x})$ y de la población G_2 con densidad $f_2(\vec{x})$, entonces, las regiones de clasificación R_1 y R_2, definidas por:

$$R_1: \quad \frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)} \geq \frac{\pi_2}{\pi_1}$$

$$R_2: \quad \frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)} < \frac{\pi_2}{\pi_1}$$

minimizan el costo total de mala clasificación. Este resultado es similar al presentado en el teorema 1 cuando los costos de mala clasificación son iguales.

Definición 3 Sean $\vec{X}_1,....,\vec{X}_n$ una muestra aleatoria de tamaño n de la población π cuya función densidad de probabilidad es $f\left(\vec{x};\vec{\theta}\right)=f\left(\vec{x}\right)$, donde $\vec{\theta}$ es un vector de parámetros. Se define la función de verosimilitud para toda la muestra por:

$L\left(\vec{\theta},\vec{X}_1,....,\vec{X}_n\right)=\prod_{i=1}^{n} f\left(\vec{x}_i,\vec{\theta}\right)$ es una función del parámetro $\vec{\theta}$ (Mardia, 1979).

Definición 4 Sean G_1 y G_2 dos poblaciones p variantes con funciones de densidad $f_1(\vec{x})$ y $f_2(\vec{x})$, la regla de clasificación basada en la máxima verosimilitud, clasifica $\vec{x}$ en la población más verosímil. Es decir, en la población que tiene mayor función de verosimilitud, o sea:

Clasificar $\vec{x}$ en G_1 cuando $L_1(\vec{x})\geq L_2(\vec{x})$, o sea cuando

$$\frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)}\geq 1 \qquad (6)$$

caso contrario, clasificar en G_2.

Observemos que la regla (6) es equivalente a la regla que minimiza CEM cuando los costos y las probabilidades a priori son iguales.

En algunas situaciones las categorías de las poblaciones se especifican de antemano, en el sentido que las distribuciones de probabilidad son conocidas. En otros casos solo se conoce la forma de las distribuciones de probabilidad por lo que, se tienen que estimar los parámetros poblacionales para lo cual se usan los resultados de las muestras de dichas poblaciones.

A continuación se presenta un resumen de la regla de clasificación basada en la función lineal discriminante de Fisher (1936), que no requiere supuestos respecto a las distribuciones de probabilidad de las poblaciones involucradas. También se presenta la regla de discriminación basada en poblaciones normales homoscedásticas y se comprobará que conducen a las mismas reglas de clasificación.

1.2.3. Regla de clasificación basada en el principio de Fisher

Fisher (1936) propuso una función discriminante para clasificar observaciones en una de dos poblaciones multivariantes, sin la suposición de normalidad y considerando implícitamente que las poblaciones son homoscedásticas. La argumentación de Fisher se basa en la transformación lineal del vector de observaciones $\vec{x}$ en una observación univariante, y, de manera que se consiga la máxima distancia o separación entre los valores , y, originados por los valores de ambas poblaciones, G_1 y G_2.

Desde el punto de vista poblacional, sean G_1 y G_2 las dos poblaciones en R^p y $\vec{X}^{T(k)} = \left(X_1^{(k)},....,\ X_p^{(k)}\right)$, $k = 1,2$, un vector aleatorio con $E\left(\vec{X}/\vec{X}\,\varepsilon\, G_1\right) = \vec{\mu}^{(1)}$ y $E\left(\vec{X}\ /\ \vec{X}\ \varepsilon\ G_2\right) = \vec{\mu}^{(2)}$ y que las matrices de covarianzas de las dos poblaciones son iguales, $\Sigma = \Sigma_1 = \Sigma_2$. Considerando una combinación lineal de las variables en estudio, $Y = \vec{\alpha}^T \vec{X}$, $\vec{\alpha}\varepsilon R^p$, las medias condicionales de la variable Y para las dos poblaciones están dadas por:

$\vec{u}_Y^{(1)} = E(Y/G_1) = \vec{\alpha}^T \vec{\mu}^{(1)}$ $\quad$ $\vec{u}_Y^{(2)} = E(Y/G_2) = \vec{\alpha}^T \vec{\mu}^{(2)}$ y la varianza de Y es

$\sigma_Y^2 = Var\left(\vec{\alpha}^T \vec{X}\right) = \vec{\alpha}^T \Sigma \vec{\alpha}$ independiente de la población de origen.

El método de Fisher consiste en obtener la combinación lineal de las coordenadas de $\vec{X}^{T(k)} = \left(X_1^{(k)},....,\ X_p^{(k)}\right)$ que maximiza , el cuadrado de las distancias entre las medias de las combinaciones lineales de los dos grupos relativo a sus varianzas, en el supuesto de que las poblaciones son homoscedásticas, es decir, $\Sigma_1 = \Sigma_2$ (Furtado, 2008).

La idea es, realizar la transformación de las variables originales pertenecientes al espacio p dimensional en las dos poblaciones, en una única dimensión, también en las dos

poblaciones. Entonces, se define una combinación lineal que maximiza la distancia estadística cuadrada entre las medias de la variable Y.

Representando esa distancia cuadrada por:

$$\Delta^2 = \frac{\left(u_Y^1 - u_Y^2\right)^2}{\sigma_Y^2} = \frac{\left(\vec{\alpha}^T \vec{\mu}^{(1)} - \vec{\alpha}^T \vec{\mu}^{(2)}\right)^2}{\sigma_Y^2}$$

$$= \frac{\vec{\alpha}^T \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right)^2}{\vec{\alpha}^T \Sigma \vec{\alpha}}$$, debemos encontrar el vector $\vec{\alpha}$ tal que

$$\max_{\vec{\alpha}} \left(\Delta^2\right) = \max_{\vec{\alpha}} \left(\frac{\left(\vec{\alpha}^T \vec{\mu}^{(1)} - \vec{\alpha}^T \vec{\mu}^{(2)}\right)^2}{\vec{\alpha}^T \Sigma \vec{\alpha}} \right)$$

Se obtiene la derivada de primer orden de Δ^2 en relación al vector $\vec{\alpha}$, igualar a cero la expresión resultante y resolver la ecuación. Después de algunas operaciones algebraicas se tiene el sistema:

$$\left(\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right)\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right)^T - \Delta^2 \Sigma \right) \vec{\alpha} = 0$$, que tendrá solución diferente de la trivial si:

$$\left| \left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)\left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)^T - \Delta^2 \Sigma \right| = 0$$

Usando propiedades de determinantes de matrices se tiene:

$$\left| \left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)\left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)^T - \Delta^2 \Sigma \right| = 1 - \frac{\left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)^T \Sigma^{-1} \left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)}{\Delta^2} = 0.$$

Por tanto,

$$\Delta^2 = \left(\vec{u}^{(1)} - \vec{u}^{(2)}\right)^T \Sigma^{-1} \left(\vec{u}^{(1)} - \vec{u}^{(2)}\right),$$

es el máximo de la distancia estadística cuadrada entre las medias de las poblaciones univariantes resultante de la transformación lineal establecida por $\vec{\alpha}$ que mejor discrimina las dos poblaciones. Ese máximo representa la distancia de Mahalanobis entre los dos centros de las dos poblaciones originales definidas en R^p (Anderson, 1984, entre otros).

El vector $\vec{\alpha}$ de las combinaciones lineales de las p variables originales, corresponde a ese máximo y es obtenido de acuerdo con las siguientes ideas:

$$\Delta^2 = \frac{\left(\vec{\alpha}^T \vec{\mu}^{(1)} - \vec{\alpha}^T \vec{\mu}^{(2)} \right)^2}{\vec{\alpha}^T \Sigma^{-1} \vec{\alpha}}$$

$$\left(\vec{\alpha}^T \vec{\mu}^{(1)} - \vec{\alpha}^T \vec{\mu}^{(2)} \right)^2 = \Delta^2 \vec{\alpha}^T \Sigma^{-1} \vec{\alpha}.$$

Si consideramos la siguiente restricción $\vec{\alpha}^T \Sigma^{-1} \vec{\alpha} = \Delta^2$, tendremos:

$$\left(\vec{\alpha}^T \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right) \right)^2 = \left(\Delta^2\right)^2$$

$$\vec{\alpha}^T \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right) = \Delta^2$$

$$\vec{\alpha}^T \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right) = \vec{\alpha}^T \Sigma^{-1} \vec{\alpha} = \Delta^2 = \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)$$

de donde se concluye que

$\vec{\alpha} = \Sigma^{-1} \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)$ Conocido como los coeficientes de la función discriminante lineal de Fisher.

Así, la combinación lineal del vector $\vec{x}$ dado por $y = \vec{\alpha}^T \vec{x} = \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \vec{x}$ es conocida como la función lineal discriminante de Fisher (Johnson,1982; Mardia, 1979; Furtado, 2008; entre otros).

La distancia estadística entre las medias de las poblaciones de las distribuciones condicionales de Y es un valor máximo dado por

$\Delta^2 = \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)$. Así, si tomáramos el punto medio entre las medias univariantes de esas distribuciones condicionales, tendremos un criterio adecuado para clasificar una observación $\vec{x}$ en una de las poblaciones.

Denominando m, ese punto medio es:

$$m = \frac{1}{2}\left(\vec{u}_Y^{(1)} + \vec{u}_Y^{(2)}\right) = \frac{1}{2}\vec{\alpha}^T \left(\vec{\mu}^{(1)} + \vec{\mu}^{(2)} \right)$$

$$= \frac{1}{2}\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} + \vec{\mu}^{(2)} \right)^2$$

Si consideramos la variable aleatoria $Y = \vec{\alpha}^T \vec{X}$, entonces:

$$E(Y/G_1) - m = E\left[\left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1} / G_1 \right] - \frac{1}{2}\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} + \vec{\mu}^{(2)} \right)$$

$$= \frac{1}{2}\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right) = \frac{1}{2}\Delta^2 .$$

Esa cantidad representa la mitad de la distancia de Mahalanobis entre los centros de las dos poblaciones, por lo que es positiva.

Así, como se trata de una esperanza matemática, tenemos la esperanza de clasificar una observación en G_1 si el resultado observado de $y - m \geq 0$. Igualmente, vamos a considerar la esperanza condicional de Y, respecto a la población G_2:

$$E(Y/G_2) - m = E\left[\left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1} / G_2 \right] - \frac{1}{2}\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} + \vec{\mu}^{(2)} \right)$$

$$= -\frac{1}{2}\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right)^T \Sigma^{-1} \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)} \right) = -\frac{1}{2}\Delta^2 .$$

El valor esperado de la diferencia o desviación de Y con respecto a m es un valor negativo, de donde se concluye que debemos clasificar $\vec{x}$ en G_2 si $y - m < 0$.

Finalmente, se establece la siguiente regla de decisión.

Clasificar $\vec{x}$ en G_1 si:

$$y \geq m \qquad (7)$$

Caso contrario clasificar en G_2.

El desarrollo de (7) conduce a la siguiente regla, denominada la regla de clasificación basada en el principio de Fisher (Lachenbruch, 1975; Furtado, 2008; Mardia, 1979; entre otros):

Clasificar $\vec{x}$ en G_1

$$y = \left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1} \vec{x} \geq \frac{1}{2}\left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1}\left(u^{(1)} + u^{(2)}\right) <==>$$
$$\left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1} \vec{x} - \frac{1}{2}\left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1}\left(u^{(1)} + u^{(2)}\right) \geq 0 \qquad (8)$$

Caso contrario clasificar en G_2.

El único supuesto sobre la función lineal discriminante de Fisher, $y = \left(u^{(1)} - u^{(2)}\right)^T \Sigma^{-1} \vec{x}$, es que las poblaciones, independiente de sus distribuciones, tienen matrices de covarianzas idénticas.

1.2.4. Regla de clasificación óptima en poblaciones normales homoscedásticas

Vamos a aplicar las reglas presentadas en (1.2.2) al caso de dos poblaciones normales multivariantes homoscedásticas.

Supongamos que la distribución del vector aleatorio $\vec{X}$, p-dimensional es $N_p\left(\vec{\mu}^{(k)};\Sigma_k\right)$, para $k=1,2$ donde $\vec{\mu}^{(k)}$ son los vectores de medias para el grupo G_k; Σ_k matrices de varianzas y covarianzas, que en estos casos asumimos $\Sigma=\Sigma_1=\Sigma_2$.

Según la regla de clasificación que minimiza el costo esperado de mala clasificación, la región de clasificación de G_1, o sea R_1, es el conjunto de los $\vec{x}$ para el cual:

$$\frac{f_1\left(\vec{x}\right)}{f_2\left(\vec{x}\right)}=\frac{(2\pi)^{-p/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\vec{x}-\vec{\mu}^{(1)}\right)^T\Sigma^{-1/2}\left(\vec{x}-\vec{\mu}^{(1)}\right)\right\}}{(2\pi)^{-p/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\vec{x}-\vec{\mu}^{(2)}\right)^T\Sigma^{-1/2}\left(\vec{x}-\vec{\mu}^{(2)}\right)\right\}}\geq\frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}$$

Tomando logaritmos y haciendo las simplificaciones algebraicas se llega a:

$$\left(u^{(1)}-u^{(2)}\right)^T\Sigma^{-1}\vec{x}-\frac{1}{2}\left(u^{(1)}-u^{(2)}\right)^T\Sigma^{-1}\left(u^{(1)}+u^{(2)}\right)\geq\ln\left(\frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}\right) \qquad (9)$$

donde el término del lado izquierdo de la desigualdad (9) es la función lineal discriminante de Fisher, ya definida antes.

Finalmente se plantea la siguiente regla de clasificación:

Asignar $\vec{x}$ al grupo G_1 si:

$$\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^T\Sigma^{-1}\vec{x}-\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^T\Sigma^{-1}\left(\vec{\mu}^{(1)}+\vec{\mu}^{(2)}\right)\geq\log\left(\frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}\right), \qquad (10)$$

caso contrario asignar $\vec{x}$ al grupo G_2.

Para costos de mala clasificación y probabilidades a priori iguales, la regla (10) es equivalente a la regla (8) obtenida usando el principio de Fisher, sin considerar el supuesto respecto a las distribuciones de probabilidad involucradas.

1.2.5. Caracterización de las probabilidades de mala clasificación

Una de las cruciales consideraciones en el análisis discriminante es la probabilidad de mala clasificación, que sirve para cuantificar el número de individuos mal clasificado mediante una determinada regla de clasificación, para poder determinarlo se tiene en cuenta las siguientes consideraciones:

Supongamos que la variable aleatoria $\vec{X}$ p-dimensional, tiene distribución $N_p\left(\vec{\mu}^{(k)};\Sigma\right)$, con k=1,2. De la combinación lineal definida anteriormente como $Y=\vec{\alpha}\vec{X}$ se puede deducir lo siguiente:

$$Y=\vec{\alpha}^{T}\vec{X}\approx N_p\left(\vec{\alpha}^{T}.\vec{\mu}^{(k)};\Delta^2\right)$$

donde: $\vec{\alpha}=\Sigma^{-1}.\left(\vec{u}^{(1)}-\vec{u}^{(2)}\right)$; $E(Y)=\vec{\alpha}^{T}.\vec{\mu}^{(k)}$ y $V(Y)=\Delta^2$

En efecto:

$$E[Y]=E\left[\vec{\alpha}^{T}.\vec{X}\right]=\vec{\alpha}^{T}E\left[\vec{X}\right]$$
$$=\vec{\alpha}^{T}.\vec{\mu}^{(k)}$$

,

$$E[Y]=\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^{T}.\Sigma.\vec{\mu}^{(k)}$$

para k=1,2, del mismo modo

$$V[Y]=V\left[\vec{\alpha}^{T}.\vec{x}\right]=\vec{\alpha}^{T}.V\left[\vec{x}\right].\vec{\alpha}$$
$$=\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^{T}.\Sigma^{-1}.\Sigma.\Sigma^{-1}\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)$$
$$=\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^{T}.\Sigma^{-1}.\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)$$
$$=\Delta^{2}$$

de este modo se tiene que $Y = \vec{\alpha}^{T}.\vec{x} \approx N\left(\vec{\alpha}^{T}.\vec{\mu}^{(k)};\Delta^2\right)$, donde

$$\Delta^2 = \left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right)^T.\Sigma^{-1}.\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right)$$

Esta última expresión conocida como la distancia de Mahalanobis, que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales, se diferencia de la distancia Euclidea porque tiene en cuenta las correlaciones de las variables aleatorias, además está considerada dentro de la familia de distancias Euclidianas ponderadas.

Las probabilidades de mala clasificación en el nuevo espacio son las siguientes, cuando $\vec{x}$ proviene del grupo G_1 entonces:

$$P(2/1) = P\left(Y > m \,/\, \text{Cuando } \vec{x} \text{ procede de } G_1\right)$$

$$P(2/1) = \left(\frac{Y - \mu_Y}{\sigma_Y} > \frac{m - \vec{\alpha}^{T}.\vec{\mu}^{(k)}}{\Delta}\right)$$

$$P(2/1) = \left(Z > \frac{\frac{1}{2}.\vec{\alpha}^{T}.\left(\vec{\mu}^{(1)} + \vec{\mu}^{(2)}\right) - \vec{\alpha}^{T}.\vec{\mu}^{(1)}}{\Delta}\right)$$

$$= 1 - \phi\left(\frac{\Delta}{2}\right)$$

donde ϕ es la distribución normal acumulada; de forma análoga, puede hallarse la probabilidad de mala clasificación, cuando $\vec{x}$ proviene del grupo G_2, entonces:

$$P(1/2) = P(Y \le m)$$

$$P(1/2) = \phi\left(-\frac{1}{2}\Delta\right)$$

Llevando en cuenta estos resultados, se puede decir, ambas probabilidades de mala clasificación, solo dependen de la distancia de Mahalanobis.

1.2.6. Estimación de parámetros y de la función lineal discriminante

En la práctica el vector de medias $\vec{\mu}^{(k)}$ y las matrices de covarianzas Σ_k son parámetros desconocidos, asociados al vector aleatorio $\vec{X}$, p variante, por lo que deben ser estimados.

Sean $\vec{X}_1, \vec{X}_2, \vec{X}_3, ..., \vec{X}_{n_k}$; $k = 1,2$ y $n_1 + n_2 = n$, muestras aleatorias del vector aleatorio $\vec{X}$ p-dimensional, que puede ser representado mediante la siguiente matriz de datos.

	Individuos	VARIABLES			
		X1	X2	.....	Xp
PRIMER GRUPO	$\vec{X}_1^{(1)T}$	$X_{11}^{(1)}$	$X_{12}^{(1)}$	.....	$X_{1p}^{(1)}$
	$\vec{X}_2^{(1)T}$	$X_{21}^{(1)}$	$X_{22}^{(1)}$	.....	$X_{2p}^{(1)}$
	.....	.....	.....	....	....
	$\vec{X}_{n_1}^{(1)T}$	$X_{n1,1}^{(1)}$	$X_{n1,2}^{(1)}$		$X_{n1,p}^{(1)}$
SEGUNDO GRUPO	$\vec{X}_1^{(2)T}$	$X_{11}^{(2)}$	$X_{12}^{(2)}$	.....	$X_{1p}^{(2)}$
	$\vec{X}_2^{(2)T}$	$X_{21}^{(2)}$	$X_{22}^{(2)}$	.....	$X_{2p}^{(2)}$
				.....	
	$\vec{X}_{n_2}^{(2)T}$	$X_{n2,1}^{(2)}$	$X_{n2,2}^{(2)}$	.....	$X_{n2,p}^{(2)}$

donde un individuo de cualquiera de los grupos es representado como: $\vec{X}_i^{(k)} = \begin{bmatrix} X_{i1}^{(k)} \\ X_{i2}^{(k)} \\ . \\ . \\ . \\ X_{ip}^{(k)} \end{bmatrix}$,

para $i = 1,2,3,...,n_k \;\; k = 1,2.$

Estimador de la media poblacional $\vec{\mu}^{(k)}$ es la media muestral $\vec{\overline{X}}^{(k)}$ definida como:

$$\vec{\overline{X}}^{(k)T} = \begin{bmatrix} \dfrac{\sum_{i=1}^{n_k} X_{i1}^{(k)}}{n_k} & \dfrac{\sum_{i=1}^{n_k} X_{i2}^{(k)}}{n_k} & \cdots & \dfrac{\sum_{i=1}^{n_k} X_{ip}^{(k)}}{n_k} \end{bmatrix}$$

$$\vec{\overline{X}}^{(k)T} = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & \overline{X}_3 & \cdots & \overline{X}_p \end{bmatrix},$$

mientras que el estimador insesgado de la matriz de varianzas y covarianzas, Σ :

$$S_u = \frac{\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right)^T}{n_1 + n_2 - 2}$$

Expresado de otra forma:

$$\begin{aligned} S_u &= \frac{(n_1 - 1).S_1 + (n_2 - 1).S_2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)}{n_1 + n_2 - 2} S_1 + \frac{(n_2 - 1)}{n_1 + n_2 - 2} S_2 \\ &= w_1.S_1 + w_2.S_2 \end{aligned}$$

donde w_1 y w_2 son los pesos de cada grupo en la formación de la matriz de varianzas covarianzas muéstrales combinada S_u.

Sustituyendo en (8) los parámetros por sus respectivos estimadores se tiene la siguiente regla muestral basada en la función lineal discriminante de Fisher:

Asignar $\vec{x}$ al l grupo G_1 si:

$$\left(\vec{\bar{X}}^{(1)} - \vec{\bar{X}}^{(2)}\right)^T S_u^{-1} \vec{x} - \frac{1}{2}\left(\vec{\bar{X}}^{(1)} - \vec{\bar{X}}^{(2)}\right)^T S_u^{-1}\left(\vec{\bar{X}}^{(1)} + \vec{\bar{X}}^{(2)}\right) \geq 0 \qquad (11)$$

caso contrario asignar al grupo G_2.

$y = \left(\vec{\bar{X}}^{(1)} - \vec{\bar{X}}^{(2)}\right)^T S_u^{-1} \vec{x}$, es el estimador de la función lineal discriminante de Fisher, presentada en (8).

Como se mencionó, las probabilidades de mala clasificación solo dependen de la distancia de Mahalanobis, cuyo estimador es expresado como:

$$D^2 = \left(\vec{\bar{X}}^{(1)} - \vec{\bar{X}}^{(2)}\right)^T S_u{}^{-1}\left(\vec{\bar{X}}^{(1)} - \vec{\bar{X}}^{(2)}\right) \qquad (12)$$

Llevando en cuenta el resultado de (12), las probabilidades de mala clasificación para ambos grupos, pueden ser estimados como:

$$P(2/1) = 1 - \phi\left(-\frac{1}{2}D\right) \quad \text{y} \quad P(1/2) = \phi\left(-\frac{1}{2}D\right)$$

Luego de construir la regla de clasificación, es necesario evaluarla a través de la probabilidad total de mala clasificación (PTM), por ejemplo. Existen varios métodos de estimación, entre ello, la razón del error aparente (Lachenbruch, 1975); la razón del error real y el estimador U modificado, propuesto por Lachenbruch y Mickey (1968). A continuación se presenta un resumen de la tasa de error aparente (TEA), por ser el método que se usará en el presente trabajo.

Existe un tipo de evaluación de la función de clasificación, a través de la estimación de las probabilidades de mala clasificación, que no depende de la forma de las poblaciones de origen y que puede calcularse para cualquier procedimiento de clasificación, denominado tasa de error aparente (Lachenbruch, 1975). Se define como la proporción de observaciones de la muestra que son mal clasificadas por la función de clasificación. Este estimador es fácil de calcular.

Se toman n_1 observaciones de G_1 y n_2 observaciones de G_2, se construye la función de clasificación de la muestra y se evalúa cada elemento $\vec{x}$ de la muestra en la función. Se obtienen la siguiente tabla:

Grupo verdadero	Decisión estadística		
	Asignar G_1	Asignar G_2	Total
G_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
G_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

donde:

n_{1c} : número de elementos de G_1 clasificados correctamente en G_1

n_{1m} : número de elementos de G_1 clasificados incorrectamente en G_2

n_{2c} : número de elementos de G_2 clasificados correctamente en G_2

n_{2m} : número de elementos de G_2 clasificados incorrectamente en G_1

Se estima la probabilidad total de mala clasificación mediante:

Tasa de error aparente (TEA) $\dfrac{n_{1m}+n_{2m}}{n_1+n_2}$

1.3. ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

1.3.1. Introducción

Análisis de regresión quizá es la metodología estadística de análisis de datos más ampliamente usada para investigar la relación entre una variable dependiente (también denominada variable respuesta) Y, y una o más variables independientes (también denominados predictores o regresores) $X_1,.....,X_P$.

Se tiene una regresión lineal múltiple cuando admitimos que la variable dependiente es función lineal, en los parámetros, de dos o más variables independientes: $X_1,.....,X_P$.

El análisis de regresión lineal múltiple tiene como objetivo principal, estimar y/o predecir, el valor medio poblacional de la variable dependiente Y, sobre la base de valores conocidos o fijos de una o más variables explicativas X_i, es decir, estimar o predecir las esperanzas o medias condicionales : $E(Y/X_i)$

El modelo de regresión lineal múltiple poblacional postula, que las medias condicionales denotadas por, $E(Y/X_i)$, son funciones lineales de X_i y pueden escribirse:

$$E(Y/X_i) = B_0 + B_1X_i \qquad (13)$$

o en su forma mas general:

$$E(Y/X_1, X_2,.....,X_P) = B_0 + B_1X_1 + B_2X_2 + ... + B_PX_P$$

donde:

$B_0, B_1,...., B_P$: son parámetros desconocidos fijos, denominados coeficiente de regresión.

B_0 recibe también el nombre de intercepto, que se obtiene cuando $X_i = 0$ es decir: $E(Y/X_i = 0) = B_0$

B_i es el coeficiente que indica el cambio en Y por cada unidad de cambio fijo en X_i.

El modelo de regresión poblacional es irreal, puesto que implica que los datos $(X;Y)$ han sido colectados sobre todos los individuos que define la población de interés. Aunque ésta no es una situación ideal, sin embargo es posible modelar de una manera muy simple:

- Si se asume que se dispone de todos los datos de la población, es posible calcular la probabilidad condicional de Y dado X_i, denotado por $P(Y / X_i)$, de la siguiente manera:

$$P(Y = Yi / X = Xj) = \frac{P(Y = Yi, X = Xj)}{P(X = Xj)}$$

con la condición $P(X = Xj) > 0$.

- La media condicional o esperanza condicional, denotado por $E(Y / X_i)$: y cuya expresión es:

$$E(Y / Xj) = \sum Yi.P(Y = Yi / X = Xj) \qquad (14)$$

es el valor medio de la probabilidad condicional de Y dado X_j.

El modelo de regresión (13) es lineal en los parámetros y en las variables, mientras que por ejemplo el modelo $E(Y / Xi) = \beta_0 + \beta_1 X_i^2$ es lineal en los parámetros pero no es lineal en las variables. Linealidad en los parámetros es la condición relevante impuesta para desarrollar la teoría que se presenta a continuación.

La diferencia $Y_i - E(Y / X_i)$, es una variable aleatoria no observable que puede tomar valores positivos o negativos y comúnmente se le conoce con el nombre de error, residuo, o término de perturbación. Así, podemos escribir:

$$Y_i = E(Y / X_i) + \varepsilon_i \qquad (15)$$

donde Y_i es igual al valor promedio de todos los valores individuales Y_i que tienen el mismo valor X_i, $E(Y / X_i)$, más una cantidad aleatoria desconocida ε_i.

- ε_i representa a todas las variables omitidas, que afecta a la variable dependiente, pero que no están incluidas en la regresión (15).
- Cabe resaltar que $Y_i \neq E(Y / X_i)$ pues, ε_i representa a todas las variables omitidas que afectan a la regresión, pero que no están incluidas en el modelo. En la práctica es posible incluir en el modelo todas las variables explicativas que se nos ocurran, pero no siempre es recomendable. Primero, porque siempre se desea tener un modelo lo más parsimonioso posible. Segundo, aún si pudiéramos cuantificar todas las otras variables "relevantes" que fueron omitidas, la influencia combinada de todas ellas puede no ser muy importante. Finalmente asumiendo que tuviésemos éxito al introducir todas las variables

"relevantes" en el modelo, siempre existe una cierta cantidad de aleatoriedad en la variable dependiente que no queda explicada por los predictores incluidos.

1.3.2. Modelo de regresión lineal múltiple

En esta sección se hace la presentación del modelo de regresión lineal múltiple con p-variantes independientes. Sin embargo, en algún momento se puede recurrir al modelo con dos variables para facilitar la exposición. En todas las situaciones en donde se use el modelo de regresión con dos variables, una independiente y otra dependiente, la generalización al modelo de regresión múltiple es directa.

Tenemos un modelo de regresión lineal múltiple cuando admitimos que el valor de la variable dependiente, Y, es función de dos o más variables independientes, $X_1,.....,X_{p-1}$ y se representa de la siguiente manera:

$$Y_i = B_0 + B_1X_1 + B_2X_2 + ... + B_{P-1}X_{p-1} + \varepsilon_i \qquad (16)$$

$$i = 1, 2, 3,....., n$$

donde:

B_0 : intercepto,

$B_0, B_1, B_2,....., B_{p-1}$: son los coeficientes de regresión parcial,

ε_i : término residual asociado a la i-ésima observación.

El modelo de regresión (16) proporciona el valor esperado de Y para valores fijos de $X_1,.....,X_{p-1}$, más una componente aleatoria.

Obsérvese que la ecuación (16) corresponde al siguiente sistema de ecuaciones:

El sistema de ecuaciones anterior escrito en forma matricial es el siguiente:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & X_{1\,p-1} \\ 1 & X_{21} & X_{22} & . & . & X_{2\,p-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & X_{np-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

$$\vec{Y}_{nx1} = X_{nxP}\vec{B}_{px1} + \vec{\varepsilon}_{nx1} \qquad (17)$$

donde :

$\vec{Y}_{nx1}$ **:** Vector de observaciones de la variable independiente,

Y_{nxp} : Matriz de resultados de n observaciones sobre $(p-1)$ variables independientes $X_1,.....,X_{p-1}$.

donde los elementos de la primera columna, unos, corresponden al intercepto B_0; esto es $X_0 = 1$.

$\vec{B}_{Px1}$: vector columna de parámetros desconocidos $B_0, B_1, B_2,...., B_{p-1}$.

B_0 es el intercepto y $B_0, B_1, B_2,...., B_{p-1}$, son denominados coeficientes de regresión parcial.

$\vec{\varepsilon}$ **:** vector columna de "n" residuos ε_i.

En el modelo de regresión lineal múltiple:

$$\vec{Y} = X\vec{B} + \vec{\varepsilon}$$

El vector $\vec{B}$ contiene $(p-1)$ coeficientes de regresión parcial, $B_0, B_1, B_2,...., B_{p-1}$, y el intercepto B_0. Se usa el término parcial porque el j-ésimo coeficiente de regresión B_j, mide el cambio en el valor medio de Y por unidad de cambio en X_j, manteniendo todas las otras variables independientes constantes. Por ejemplo en un modelo de regresión con X_1, X_2, como variables independientes, B_2 mide el cambio en el valor de, $E(Y / X_1, X_2)$, con respecto a X_2, manteniendo constante la influencia de X_1.

Supuestos del modelo:

Como ya se dijo, los estimadores obtenidos por el método de Mínimos Cuadrados Ordinarios poseen propiedades muy deseables. Sin embargo, se sustenta en los siguientes supuestos:

1. **S1:** $E\left(\vec{\varepsilon}\right) = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ .\\ \varepsilon_n\end{pmatrix} = \begin{pmatrix}0\\ 0\\ .\\ 0\end{pmatrix} = \vec{0}$

2. **S2:** $E(\vec{\varepsilon}) = Var(\vec{\varepsilon}) = \delta^2 I$

$E(\varepsilon_i\, \varepsilon_j) = 0 \qquad \forall i \neq j$

$E(\varepsilon_i\, \varepsilon_j) = E(\varepsilon_i^2) = \delta^2 \qquad \forall i = j$

$E(\varepsilon_1^2) \qquad E(\varepsilon_1\, \varepsilon_2) \quad \ldots\ldots \quad E(\varepsilon_1\, \varepsilon_n)$ –

en otras palabras:

El valor esperado de los residuos es cero;

No existe correlación entre el i-ésimo y el j-ésimo residuo;

La varianza de los residuos es constante.

3. **S3:** $Cov\left(\vec{\varepsilon}, X\right) = 0$, es decir, la covarianza entre las X_j y el término residual $\vec{\varepsilon}$ es cero. Este supuesto es obvio si las variables X no son aleatorias, de manera que la matriz X está formada por números fijos.

4. S4: $r(X) = p$ donde p<n,

es decir, el rango de la matriz X es "p", igual al numero de columnas de X, siempre menor que " n" el número de observaciones. La suposición 4 es más conocida como " no existe multicolinealidad ".

Cuando son satisfechas las condiciones S1 , S2 , S3 , S4 , entonces los estimadores obtenidos por el método de Mínimos Cuadrados Ordinarios (OLS), pertenecen a la clase de estimadores insesgados de mínima varianza, (BLUE) , Esto es, son los mejores estimadores lineales insesgados .

La propiedad de insesgado unida a la de mínima varianza significa que los estimadores obtenidos por el Método de Mínimos Cuadrados Ordinarios (OLS) son también estimadores eficientes.

Cuando necesitamos hacer pruebas de hipótesis y obtener intervalos de confianza hay la necesidad de imputar una distribución de probabilidades para los errores, por tanto, tenemos el siguiente supuesto adicional a los ya presentados.

5. S5: los errores tienen distribución normal multivariante.

$$\vec{\varepsilon} \sim N(0, \delta^2 I) \qquad E(\vec{\varepsilon}) = 0 \qquad V(\vec{\varepsilon}) = \delta^2 I$$

Usando los supuestos anteriores, el Método de Mínimos Cuadrados Ordinarios (OLS) nos permite estimar los parámetros $\vec{B}$ del modelo:

$$\vec{Y} = X\vec{B} + \vec{\varepsilon}$$

Es decir, debemos encontrar $\vec{B}$ tal que:

$$Q(\vec{B}) = |\vec{Y} - \mathbf{X}\vec{B}|^2 = \min|\vec{Y} - \mathbf{X}\vec{B}|^2 = \min \vec{\varepsilon}\,\vec{\varepsilon}' = \min(\vec{Y} - \mathbf{X}\vec{B})'(\vec{Y} - \mathbf{X}\vec{B})$$

$$= \min(\vec{Y}'\vec{Y} - \mathbf{2}\vec{B}'\mathbf{X}'Y + \vec{B}'\mathbf{X}'\mathbf{X}\vec{B}).$$

Para obtener $\vec{B}$ se resuelve: $\dfrac{\partial Q(\vec{B})}{\partial \vec{B}} = 0$

donde $Q(\vec{B}) = |\vec{Y} - \mathbf{X}\vec{B}|^2$.

La solución es

$\vec{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y}$, donde $\vec{B}$ es el estimador OLS de $\vec{B}$.

1.4. RELACIÓN ENTRE EL ANÁLISIS DE DISCRIMINANTE LINEAL Y EL ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

1.4.1. Introducción

Dado el objetivo de la tesis, es la identificación de observaciones influyentes en el marco del análisis discriminante, es de particular importancia establecer un paralelo, en lo posible, entre las dos metodologías estadísticas, el análisis discriminante lineal en dos grupos y el análisis de regresión lineal múltiple. Se demostrará que a pesar de ser dos metodologías filosóficamente diferentes, computacionalmente, los coeficientes, de la función lineal discriminante de Fisher y de la función de regresión lineal múltiple, guardan una proporcionalidad (Anderson, 1984). Cabe indicar que el resultado presentado en resumen en la referencia citada, en el presente trabajo, se desarrolla en detalle los aspectos teóricos involucrados.

En el siguiente cuadro se presenta el paralelo entre ambos métodos estadísticos.

Análisis discriminante lineal	Análisis de regresión lineal múltiple
La variable dependiente es binaria.	La variable dependiente es cuantitativa. Para efectos de pruebas de hipótesis, se supone que la variable dependiente tiene distribución normal.
Las variables independientes son aleatorias. Para efectos inferenciales se supone que dichas variables tienen distribución normal.	Las variables independientes son supuestas fijas.
El objetivo es encontrar la combinación lineal de las variables independientes que permiten establecer una regla de clasificación con el fin de minimizar la probabilidad total de mala clasificación.	Su objetivo es predecir el valor E(Y/Xi)
Es una estrategia para encontrar una manera de clasificar individuos.	Es un modelo formal con suposiciones para estimar parámetros.

A continuación se presenta en detalle el desarrollo teórico y metodológico que permite encontrar la relación que existe entre los coeficientes del la función lineal discriminante de Fisher y los coeficientes del modelo de regresión lineal múltiple.

1.4.2. Metodología que establece la relación entre coeficientes

Supongamos que se han tomado muestras aleatorias $\vec{X}_i^{(k)}$ $i = 1,2,3,...,n_k$ $k = 1,2$ de tamaños n_1 y n_2 de cada uno de los grupos y que puede expresarse mediante la matriz, presentada en la sección anterior.

Se definen, $Y_i^{(1)}$ y $Y_i^{(2)}$ que emulan los valores de la variable dependiente (binaria) en cada uno de los grupos; en lugar de los códigos, 1 y 2, que tradicionalmente se usan para designar a los grupos involucrados en el análisis discriminante. Toman el mismo valor en cada uno de los grupos, según la siguiente fórmula:

$$Y_i^{(1)} = \frac{n_2}{n_1 + n_2} \quad \text{y} \quad Y_i^{(2)} = -\frac{n_1}{n_1 + n_2}$$

Se ajusta el modelo de regresión lineal múltiple, con $Y_i^{(1)}$ y $Y_i^{(2)}$ como los valores de la variable dependiente binaria y los valores de $\vec{X}_i^{k}$ como los de las variables explicativas, es decir:

$$Y_i^{(k)} = \beta_0 + \beta_{11} X_{i1}^{(k)} + \beta_2 X_{i1}^{(k)} + ... + \beta_p X_{ip}^{(k)} + \vec{\varepsilon}_i^{(k)}$$

donde $i = 1,2,...,n_k$ y $k = 1,2$

Mediante el método de los mínimos cuadrados se procede a estimar los parámetros involucrados, es decir, minimizar:

$$SCE = \sum_{K=1}^{2} \sum_{I=1}^{n_g} \left[\varepsilon_i^{(k)}\right]^2 = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left[Y_i^{(k)} = -\beta_0 - \beta_1 X_{i1}^{(k)} - \beta_2 X_{i1}^{(k)} - ... - \beta_p X_{ip}^{(k)}\right]^2$$

Para facilitar los cálculos se replantea y se resuelve con datos centrados, para ello se realizan las siguientes cuentas previas:

$$X^* = \vec{X} - \vec{\overline{X}} \qquad ; \qquad \vec{\overline{X}} = \frac{n_1 . \vec{\overline{X}}^{(1)} + n_1 . \vec{\overline{X}}^{(2)}}{n_1 + n_2}$$

y se debe resolver el siguiente sistema.

$$\underbrace{\vec{Y}}_{[(n_1+n_2)x1]} = \underbrace{\vec{\beta}}_{[(n_1+n_2)xp]} . \underbrace{X^{*}}_{[px1]} + \underbrace{\vec{\varepsilon}}_{[(n_1+n_2)x1]} \qquad (18)$$

A continuación se presenta paso a paso la metodología que conduce a la estimación del vector de parámetros involucrado en (18).

Se forman las siguientes ecuaciones normales:

$$\left[\vec{X}_1^{(1)} - \vec{\bar{X}}, \vec{X}_2^{(1)} - \vec{\bar{X}}, ..., \vec{X}_{n_1}^{(1)} - \vec{\bar{X}}, \vec{X}_1^{(2)} - \vec{\bar{X}}, \vec{X}_2^{(2)} - \vec{\bar{X}}, ..., \vec{X}_{n_2}^{(2)} - \vec{\bar{X}}\right] . \begin{bmatrix} \left(\vec{X}_1^{(1)} - \vec{\bar{X}}\right)^T \\ . \\ . \\ \left(\vec{X}_{n_1}^{(1)} - \vec{\bar{X}}\right)^T \\ \left(\vec{X}_1^{(2)} - \vec{\bar{X}}\right)^T \\ . \\ . \\ \left(\vec{X}_{n_2}^{(2)} - \vec{\bar{X}}\right)^T \end{bmatrix} . \vec{\beta} = . \left[\vec{X}_1^{(1)} - \vec{\bar{X}}, \vec{X}_2^{(1)} - \vec{\bar{X}}, ..., \vec{X}_{n_1}^{(1)} - \vec{\bar{X}}, \vec{X}_1^{(2)} - \vec{\bar{X}}, \vec{X}_2^{(2)} - \vec{\bar{X}}, ..., \vec{X}_{n_2}^{(2)} - \vec{\bar{X}}\right] . \vec{Y}$$

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k} \left(\vec{X}_i^{(k)} - \vec{\bar{X}}\right) . \left(\vec{X}_i^{(k)} - \vec{\bar{X}}\right)^T \vec{\hat{\beta}} = \sum_{k=1}^{2}\sum_{i=1}^{n_k} \left(\vec{X}_i^{(k)} - \vec{\bar{X}}\right) . Y_i^{(k)}$$

Agregando y quitando el término $\vec{\bar{X}}^{(k)}$ en el lado izquierdo de la expresión anterior, además desarrollando la primera sumatoria de la expresión que está a lado derecho.

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k} \left(\left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)\right) . \left(\left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)\right)^T . \vec{\hat{\beta}} = \sum_{i=1}^{n_1} \left(\vec{X}_i^{(1)} - \vec{\bar{X}}\right) . Y_i^{(1)} + \sum_{i=1}^{n_2} \left(\vec{X}_i^{(2)} - \vec{\bar{X}}\right) . Y_i^{(2)}$$

$$\underbrace{\sum_{k=1}^{2}\sum_{i=1}^{n_k} \left(\left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)\right) . \left(\left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)\right)^T}_{F} . \vec{\hat{\beta}} = \underbrace{\sum_{i=1}^{n_1} \left(\vec{X}_i^{(1)} - \vec{\bar{X}}\right) . \frac{n_1}{n_1+n_2} - \sum_{i=1}^{n_2} \left(\vec{X}_i^{(2)} - \vec{\bar{X}}\right) . \frac{n_1}{n_1+n_2}}_{G} \quad ...(19)$$

Para simplificar los cálculos, la última expresión se divide en dos expresiones, aplicando propiedades de transpuesta y multiplicación de matrices, solo en la expresión F, se tiene:

$$F = \sum_{k=1}^{2}\sum_{i=1}^{n_k} \left[\left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) . \left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right)^T + \left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right) . \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)^T + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right) . \left(\vec{X}_i^{(k)} - \vec{\bar{X}}^{(k)}\right)^T + \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right) . \left(\vec{\bar{X}}^{(k)} - \vec{\bar{X}}\right)^T \right]$$

Distribuyendo las sumatorias:

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right)^T+\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right)^T$$

De la última expresión se demuestra que (véase en el apéndice A):

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right)^T=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T=0$$

Por lo que esta, queda expresada como:

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(k)}-\vec{\overline{X}}\right)^T$$

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\sum_{i=1}^{n_1}\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}\right)^T+\sum_{i=1}^{n_2}\left(\vec{\overline{X}}^{(2)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(2)}-\vec{\overline{X}}\right)^T$$

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+n_1.\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}\right)^T+n_2\left(\vec{\overline{X}}^{(2)}-\vec{\overline{X}}\right).\left(\vec{\overline{X}}^{(2)}-\vec{\overline{X}}\right)^T$$

Reemplazando el valor de $\vec{\overline{X}}$ definida anteriormente, se tiene:

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+n_1.\left(\vec{\overline{X}}^{(1)}-\frac{n_1.\vec{\overline{X}}^{(1)}+n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right).\left(\vec{\overline{X}}^{(1)}-\frac{n_1.\vec{\overline{X}}^{(1)}+n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right)^T$$
$$+n_2\left(\vec{\overline{X}}^{(2)}-\frac{n_1.\vec{\overline{X}}^{(1)}+n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right).\left(\vec{\overline{X}}^{(2)}-\frac{n_1.\vec{\overline{X}}^{(1)}+n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right)^T$$

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+n_1.\left(\frac{n_2.\vec{\overline{X}}^{(1)}-n_2.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right).\left(\frac{n_2.\vec{\overline{X}}^{(1)}-n_2.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right)^T$$
$$+n_2\left(\frac{n_1.\vec{\overline{X}}^{(1)}-n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right).\left(\frac{n_1.\vec{\overline{X}}^{(1)}-n_1.\vec{\overline{X}}^{(2)}}{n_1+n_2}\right)^T$$

$$F=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\frac{n_1.n_2}{n_1+n_2}\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T$$

En la expresión G, definida en (19)

$$G=\sum_{i=1}^{n_1}\left(\vec{X}_i^{(1)}-\vec{\overline{X}}\right).\frac{n_1}{n_1+n_2}-\sum_{i=1}^{n_2}\left(\vec{X}_i^{(2)}-\vec{\overline{X}}\right).\frac{n_1}{n_1+n_2}$$

$$G=\frac{n_1}{n_1+n_2}.\left[\sum_{i=1}^{n_1}\vec{X}_i^{(1)}-\sum_{i=1}^{n_1}\vec{\overline{X}}\right]-\frac{n_1}{n_1+n_2}.\left[\sum_{i=1}^{n_2}\vec{X}_i^{(2)}-\sum_{i=1}^{n_2}\vec{\overline{X}}\right]$$

$$G=\frac{n_1}{n_1+n_2}.\left[n_1.\vec{\overline{X}}^{(1)}-n_1.\vec{\overline{X}}\right]-\frac{n_1}{n_1+n_2}.\left[n_2.\vec{\overline{X}}^{(2)}-n_2.\vec{\overline{X}}\right]$$

$$G=\frac{n_1.n_2}{n_1+n_2}.\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]$$

Reemplazando los resultados obtenidos de F y G en (19)

$$\left[\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T+\frac{n_1.n_2}{n_1+n_2}\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T\right].\vec{\beta}=\frac{n_1.n_2}{n_1+n_2}.\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]$$

Sabemos que $\dfrac{\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right).\left(\vec{X}_i^{(k)}-\vec{\overline{X}}^{(k)}\right)^T}{n_1+n_2-2}=S_u$, entonces la expresión anterior toma la siguiente forma:

$$\left[(n_1+n_2-2)..S_u+\frac{n_1.n_2}{n_1+n_2}\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T\right].\vec{\beta}=\frac{n_1.n_2}{n_1+n_2}.\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]$$

$$(n_1+n_2-2).S_u.\vec{\beta}+\frac{n_1.n_2}{n_1+n_2}\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T\vec{\beta}=\frac{n_1.n_2}{n_1+n_2}.\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]$$

Tomado Transpuesta, a ambos miembros de esta última igualdad, se tiene:

$$(n_1+n_2-2)\vec{\beta}^T\ S_u+\frac{n_1.n_2}{n_1+n_2}\hat{\vec{\beta}}^T\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)=\frac{n_1.n_2}{n_1+n_2}.\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]^T$$

Introduciendo S_u^{-1} (S_u es definida positiva y no singular) se tiene:

$$(n_1+n_2-2)\vec{\beta}^T\ S_u^{-1}S_u+\frac{n_1.n_2}{n_1+n_2}\hat{\vec{\beta}}^T\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)^T S_u^{-1}\left(\vec{\overline{X}}^{(1)}-.\vec{\overline{X}}^{(2)}\right)=\frac{n_1.n_2}{n_1+n_2}S_u^{-1}\left[\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right]^T$$

Teniendo en cuenta que S_u^{-1} es simétrica, por lo que la transpuesta de dicha matriz es la misma,

entonces:

$$(n_1+n_2-2)\hat{\vec{\beta}}^{T}+\left[\frac{n_1.n_2}{n_1+n_2}D^2\right].\hat{\vec{\beta}}^{T}=\frac{n_1.n_2}{n_1+n_2}\left[S_u^{-1}\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right)\right]^T$$

$$(n_1+n_2-2)\hat{\vec{\beta}}^{T}+\frac{n_1.n_2}{n_1+n_2}D^2\hat{\vec{\beta}}^{T}=\frac{n_1.n_2}{n_1+n_2}\hat{\vec{\beta}}^{T}$$

$$\hat{\vec{\beta}}^{T}.\left[(n_1+n_2-2)+\frac{n_1.n_2}{n_1+n_2}D^2\right]=\frac{n_1.n_2}{n_1+n_2}.\hat{\vec{\alpha}}^{T}\quad,\quad\text{donde}\quad\hat{\alpha}^{T}=S_u^{-1}\left(\vec{\overline{X}}^{(1)}-\vec{\overline{X}}^{(2)}\right)\ \text{es el}$$

estimador de los coeficientes de la función lineal discriminante de Fisher.

Finalmente se tiene:

$$\hat{\vec{\beta}}^{T}=\frac{\frac{n_1.n_2}{n_1+n_2}}{(n_1+n_2-2)+\frac{n_1.n_2}{n_1+n_2}.D^2}.\hat{\vec{\alpha}}^{T}$$

$$\hat{\vec{\beta}}^{T}=K.\hat{\vec{\alpha}}^{T}\qquad(20)$$

donde: $$K=\frac{\frac{n_1.n_2}{n_1+n_2}}{(n_1+n_2-2)+\frac{n_1.n_2}{n_1+n_2}.D^2}$$

El resultado presentado en la ecuación (20), demuestra que existe una relación de proporcionalidad entre los coeficientes de la función lineal discriminante de Fisher, involucrado en el problema de discriminación y clasificación lineal y, los coeficientes de la recta de regresión lineal múltiple. Asimismo, en (1.4.1) se ha establecido un paralelo entre ambas metodologías estadísticas. Se puede afirmar que, a pesar de ser dos metodologías filosóficamente diferentes y con objetivos y supuestos diferentes, computacionalmente, solo son diferentes por la constante de proporcionalidad, k.

El resultado (20) nos plantea retos y la siguiente interrogante: Será posible usar las medidas de influencia ampliamente utilizadas en el análisis de regresión lineal múltiple, para detectar observaciones influyentes en el análisis discriminante lineal.

1.4.3. Algunas aplicaciones de los resultados teóricos

Para ilustrar los resultados teóricos que muestran la relación existente entre los coeficientes de la función lineal discriminante de Fisher y los coeficientes de la recta de regresión lineal múltiple, se consideran tres conjunto de datos.

El primer conjunto de datos corresponden a una muestra de gorriones moribundos, algunos de los cuales murieron mientras que otros sobrevivieron a la tormenta, que fueron estudiados en el laboratorio de Brown University, Rhode Island en febrero de 1898. Cuando Bumpus recolectó los datos, su interés principal fue postular o lanzar algunas propuestas a la Teoría de Darwin sobre selección natural. El concluyó a partir de los datos que: " las aves que mueren (no por razones de accidente) es porque están físicamente descalificadas y las que sobreviven lo hacen porque poseen ciertas características físicas que las ayudan a sobrevivir. Las variables estudiadas fueron: Longitud total, longitud del ala, longitud del pico y cabeza, longitud del húmero y longitud del Keel del externón(Manly, 2005).

El Segundo conjunto de datos, corresponde a una investigación conducida los años 2004 y 2005, por un grupo de investigadores del departamento de Etnobotánica y Botánica Económica del Museo de Historia Natural y de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos. De dicho estudio se consideran solo tres de las mediciones de 100 plantas medicinales del género *Minthostachys*: Longitud del peciolo, largo de la hoja y ancho de la hoja, en plantas con escasa y abundante pubescencia (Gómez, et. al., 2008).

El tercer conjunto de datos es tradicionalmente usado cuando se trata de estudiar el problema de clasificación en el análisis multivariante(Fisher, 1936). Son 4 las variables estudiadas, largo y ancho del sépalo, largo y ancho del pétalo, en tres especies de iris: setosa, versicolor y virginica.

a. Primer conjunto de datos

El primer conjunto de datos corresponden a una muestra de gorriones moribundos(algunos de lso cuales murieron mientras que otros sobrevivieron) que fueron estudiados en un laboratorio de Brown University, Rhode Island en febrero de 1898. Cuando Bumpus recolectó los datos, su interés principal fue postular o lanzar algunas propuestas a la Teoría de Darwin sobre selección natural. El concluyó a partir de los datos que: " las aves que mueren (no por razones de accidente) es porque están físicamente descalificadas y las que sobreviven lo hacen porque poseen ciertas características físicas que las ayudan a sobrevivir(Manly, 2005).

El primer conjunto de datos está conformado por:

Grupo G_1: 21 gorriones que sobrevivieron a la tormenta

Grupo G_2: 28 gorriones que no sobrevivieron a la tormenta

Las variables estudiadas en ambos grupos fueron:

X_1 = Longitud total del gorrión

X_2 = Longitud del ala

X_3 = Longitud del pico y cabeza

X_4 = Longitud del húmero

X_5 = Longitud del Keel del externón.

a.1. Resultados involucrados en el análisis del análisis discriminante lineal

Vectores de medias y matrices de covarianzas de los gorriones sobrevivientes y de los no sobrevivientes:

$$\vec{\overline{X}}^{(1)} = \begin{bmatrix} 157.380 \\ 241.000 \\ 31.4330 \\ 18.5000 \\ 20.8095 \end{bmatrix} \qquad \vec{\overline{X}}^{(2)} = \begin{bmatrix} 158.4286 \\ 241.5714 \\ 3.478600 \\ 18.44640 \\ 20.83930 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 11.05 & 9.10 & 1.56 & 0.87 & 1.29 \\ 9.10 & 17.5 & 1.91 & 1.31 & 0.88 \\ 1.56 & 1.91 & 0.53 & 0.19 & 0.24 \\ 0.87 & 1.31 & 0.19 & 0.18 & 0.13 \\ 1.28 & 0.88 & 0.24 & 0.13 & 0.58 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 11.05 & 9.10 & 1.56 & 0.87 & 1.29 \\ 9.10 & 17.50 & 1.91 & 1.31 & 0.88 \\ 1.56 & 1.91 & 0.53 & 0.19 & 0.24 \\ 0.87 & 1.31 & 0.19 & 0.18 & 0.13 \\ 1.29 & 0.88 & 0.24 & 0.13 & 0.58 \end{bmatrix}$$

Matriz de covarianzas combinada:

$$S_u = \begin{bmatrix} 13.36 & 13.75 & 1.95 & 1.37 & 2.23 \\ 13.75 & 26.15 & 2.76 & 2.25 & 2.71 \\ 1.95 & 2.76 & 0.64 & 0.35 & 0.42 \\ 1.37 & 2.25 & 0.35 & 0.33 & 0.35 \\ 0.23 & 2.71 & 0.42 & 0.35 & 1.00 \end{bmatrix}$$

La prueba de hipótesis, para contrastar la hipótesis de igualdad de matrices de covarianzas, arrojó el siguiente valor para el estadístico M de Box(a través del software SPSS):

-2Log(M_Box)	10.952
p-valor	0.841
Nivel de significación	0.05

Como el valor de la probabilidad asociada a la estadística M de Box, p-valor=0.841, es mayor que el nivel de significación prefijado,0.05, no se rechaza la hipótesis de que las matrices de covarianzas son iguales.

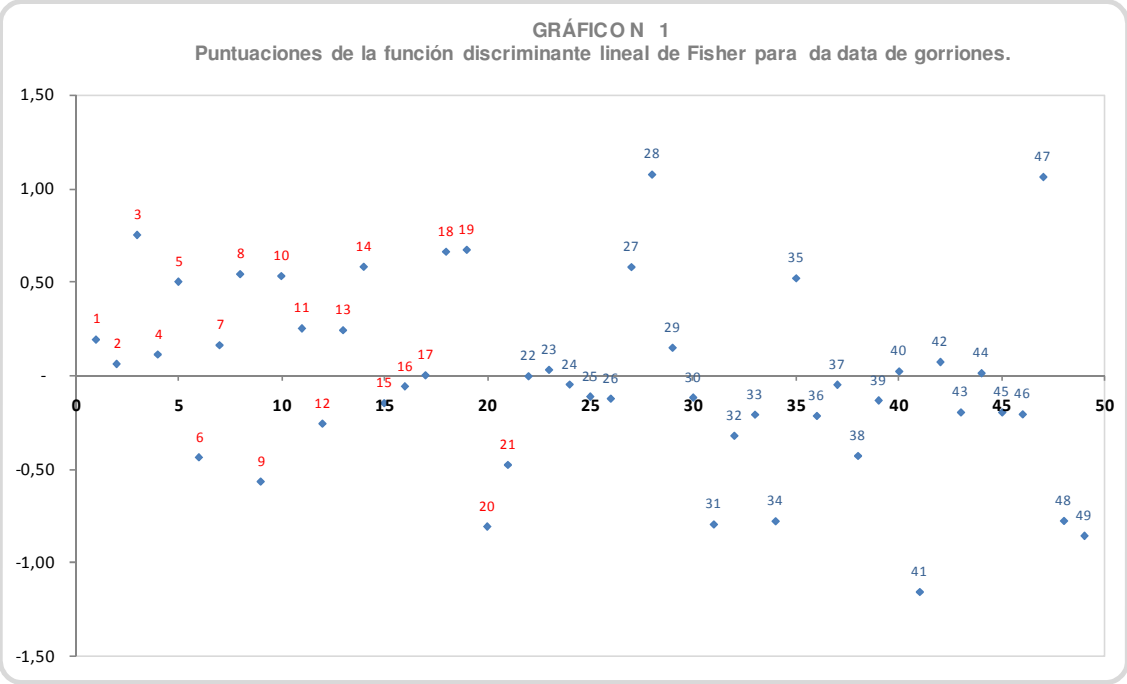
La estimación del vector de coeficientes de la función lineal discriminante de Fisher, presentadas en (1.2.3), fue:

$$\left[\hat{\alpha}\right] = \begin{bmatrix} -0.1553 \\ -0.0265 \\ -0.0929 \\ 1.0325 \\ 0.0693 \end{bmatrix}$$

Las puntuaciones y la representación gráfica de la función discriminante de Fisher, se muestran en la TABLA Nº 1 el GRÀFICO Nº 1 respectivamente.

TABLA N° 1 Puntuaciones de la función discriminante lineal de Fisher, para la data de gorriones									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
1	0,19	11	0,25	21	-0,48	31	-0,80	41	-1,16
2	0,06	12	-0,26	22	-0,01	32	-0,32	42	0,07
3	0,75	13	0,24	23	0,03	33	-0,21	43	-0,20
4	0,11	14	0,58	24	-0,05	34	-0,78	44	0,01
5	0,50	15	-0,15	25	-0,12	35	0,52	45	-0,20
6	-0,44	16	-0,06	26	-0,13	36	-0,22	46	-0,21
7	0,16	17	-0,00	27	0,58	37	-0,05	47	1,06
8	0,54	18	0,66	28	1,07	38	-0,43	48	-0,78
9	-0,57	19	0,67	29	0,15	39	-0,14	49	-0,86
10	0,53	20	-0,81	30	-0,12	40	0,02		

Las observaciones mal clasificados en el primer grupo son: 6, 9, 12, 15, 16, 17, 20, 21, y en el segundo grupo: 23, 27, 28, 29, 35, 40, 42, 44, 47.



En el siguiente cuadro se presenta los resultados de la clasificación, con cuyos resultados se estimó la capacidad predictiva del modelo

Gorriones	Decisión estadística		
	sobrevivieron	No sobrevivieron	Total
sobrevivieron	13	8	21
No sobrevivieron	9	19	28
	22	27	49

La $\text{tasa de error aparente}(\text{TEA}) = \frac{8+9}{49}(100) = 34.7\%$

Esto es, usando la función lineal discriminante de Fisher, solo el 65.3% de los gorriones fueron bien clasificadas. A pesar de cumplir con el supuesto de homocedasticidad de las matrices de covarianzas la capacidad predictiva de la función discriminante lineal de Fisher puede considerarse como pobre.

a.2. Resultados involucrados en el análisis de regresión lineal múltiple

Vectores de medias de cada grupo y el vector de medias combinado:

$$\overrightarrow{\bar{X}}^{(1)} = \begin{bmatrix} 157.380 \\ 241.000 \\ 31.4330 \\ 18.5000 \\ 20.8095 \end{bmatrix} \qquad \overrightarrow{\bar{X}}^{(2)} = \begin{bmatrix} 158.4286 \\ 241.5714 \\ 3.478600 \\ 18.44640 \\ 20.83930 \end{bmatrix}$$

$$\overrightarrow{\bar{X}} = \begin{bmatrix} 157.98 \\ 241.33 \\ 31.460 \\ 18.470 \\ 20.830 \end{bmatrix}$$

Matriz de datos centrados de las variables explicativas:

Que consiste en quitarle a cada observación el vector de medias combinado, obtenido en el paso anterior, Luego, dichos datos, que se encuentran en el anexo B, se usan como valores de las 5 variables explicativas.

El vector de datos de la variable dependiente(para cada grupo) se genera mediante la siguiente fórmula::

$$Y_i^{(1)} = \frac{n_2}{n_1 + n_2} = \frac{21}{49} = 0.43 \qquad \text{y} \qquad Y_i^{(2)} = -\frac{n_1}{n_1 + n_2} = \frac{-28}{49} = -0.57,$$

Los 21 valores de la variable dependiente, correspondiente al grupo de gorriones sobrevivientes, toman el valor 0.43 y los 28 valores de la variable dependiente, correspondiente al grupo de gorriones que no sobrevivieron, toman el valor -0.57.

Así, el vector resultante de coeficientes de la recta de regresión lineal múltiple es:

$$\hat{\beta} = \begin{bmatrix} 0.0052 \\ 0.0009 \\ 0.0031 \\ -0.0348 \\ -0.0023 \end{bmatrix}.$$

Con lo cual, se verifica que $\hat{\beta} = k\hat{\alpha}$, es decir,

$$\hat{\beta} = \begin{bmatrix} 0.0052 \\ 0.0009 \\ 0.0031 \\ -0.0348 \\ -0.0023 \end{bmatrix} = k\left[\hat{\alpha}\right] = -\frac{1}{29.6570}\begin{bmatrix} -0.1553 \\ -0.0265 \\ -0.0929 \\ 1.0325 \\ 0.0693 \end{bmatrix},$$

donde $k = -\frac{1}{29.6570}$.

Se ha comprobado computacionalmente que existe una relación de proporcionalidad entre los coeficientes de la función lineal discriminante de Fisher y los coeficientes de la recta de regresión lineal múltiple.

b. Segundo conjunto de datos

El Segundo conjunto de datos, corresponde a una investigación conducida los años 2004 y 2005, por investigadores del Departamento de Etnobotánica y Botánica Económica del Museo de Historia Natural y de la Facultad de Ciencias Biológicas de la Universidad Nacional Mayor de San Marcos. De dicho estudio se han tomado tres mediciones de 100 plantas medicinales del género *Minthostachys* (Gómez, et. al., 2008), en ella han considerado lo siguiente:

Grupo G_1: 51 plantas de *Minthostachys* con abundante pubescencia
Grupo G_2: 49 plantas de *Minthostachys* con escasa pubescencia
Las variables estudiadas en ambos grupos fueron:
X_1 = Longitud el peciolo

X_2 = Largo de la hoja

X_3 = Ancho de la Hoja

b.1. Resultados involucrados en el análisis del análisis discriminante lineal

Vectores de medias y matrices de covarianzas de las minthostachys con abundante y escasa pubescencia son:

$$\overrightarrow{\overline{X}}^{(1)} = \begin{bmatrix} 0.475 \\ 3.243 \\ 1.736 \end{bmatrix} \qquad \overrightarrow{\overline{X}}^{(2)} = \begin{bmatrix} 1.202 \\ 3.671 \\ 2.167 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 0.016 & 0.017 & 0.015 \\ 0.017 & 0.304 & 0.061 \\ 0.015 & 0.061 & 0.122 \end{bmatrix} \quad S_2 = \begin{bmatrix} 0.239 & 0.259 & 0.186 \\ 0.259 & 0.592 & 0.344 \\ 0.186 & 0.344 & 0.275 \end{bmatrix}$$

Matriz de covarianzas combinada:

$$S_u = \begin{bmatrix} 0.1252 & 0.1352 & 0.0989 \\ 0.1352 & 0.4449 & 0.1998 \\ 0.0989 & 0.1998 & 0.1971 \end{bmatrix}$$

La prueba de hipótesis, para contrastar la hipótesis de igualdad de matrices de covarianzas, arrojó el siguiente valor para el estadístico M de Box(a través del software SPSS):

-2Log(M_Box)	83.513
p-valor	0.0001
Nivel de significación	0.05

Como el valor de la probabilidad asociada a la estadística M de Box , p-valor= 0.0001, es menor que el nivel de significación prefijado, 0.05, se rechaza la hipótesis de que las matrices de

covarianzas son iguales. Al nivel de significación 0.05, las matrices de varianzas y covarianzas son diferentes.

El vector de coeficientes de la función lineal discriminante de Fisher(1.2.3), fue el siguiente:

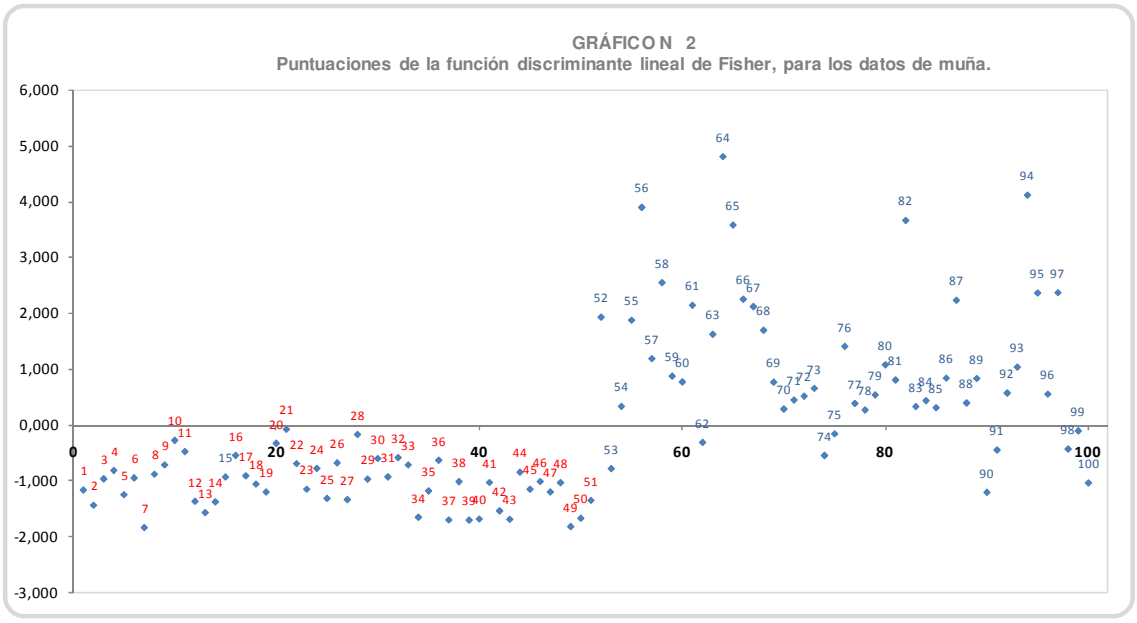
$$\left[\hat{\alpha}\right] = \begin{bmatrix} -7.2491 \\ 1.0795 \\ 0.3566 \end{bmatrix}$$

Las puntuaciones de la función discriminante lineal de Fisher se muestran a en la Tabla Nº 2 y la representación gráfica de dichas puntuaciones en el Gráfico Nº 2.

TABLA N° 2
Puntuaciones de la función discriminante lineal de Fisher, para la data de muña

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
1	-1,16	21	-0,08	41	-1,03	61	2,15	81	0,81
2	-1,44	22	-0,69	42	-1,54	62	-0,31	82	3,67
3	-0,96	23	-1,15	43	-1,69	63	1,63	83	0,33
4	-0,81	24	-0,78	44	-0,84	64	4,81	84	0,44
5	-1,25	25	-1,31	45	-1,15	65	3,58	85	0,31
6	-0,95	26	-0,68	46	-1,01	66	2,25	86	0,84
7	-1,84	27	-1,34	47	-1,20	67	2,12	87	2,23
8	-0,88	28	-0,17	48	-1,03	68	1,70	88	0,40
9	-0,71	29	-0,97	49	-1,82	69	0,77	89	0,83
10	-0,27	30	-0,60	50	-1,67	70	0,29	90	-1,21
11	-0,47	31	-0,93	51	-1,35	71	0,45	91	-0,45
12	-1,37	32	-0,58	52	1,93	72	0,52	92	0,57
13	-1,57	33	-0,71	53	-0,78	73	0,66	93	1,04
14	-1,38	34	-1,65	54	0,33	74	-0,54	94	4,12
15	-0,93	35	-1,18	55	1,88	75	-0,16	95	2,37
16	-0,54	36	-0,63	56	3,90	76	1,41	96	0,55
17	-0,91	37	-1,70	57	1,19	77	0,39	97	2,37
18	-1,06	38	-1,01	58	2,55	78	0,27	98	-0,43
19	-1,20	39	-1,70	59	0,88	79	0,54	99	-0,10
20	-0,33	40	-1,68	60	0,77	80	1,08	100	-1,04

Las observaciones que fueron mal clasificadas, todas pertenecen al segundo grupo: 53, 62, 74, 75, 90, 91, 98, 99 y la 100.



También se obtuvo la capacidad predictiva de la función discriminante, usando los resultados de la clasificación que se presenta en la siguiente tabla.

Especie Verdadera	Decisión estadística		
	Pubescencia abundante	Pubescencia escasa	Total
Pubescencia abundante	51	0	51
Escasa Pubescencia	9	40	49

$$\text{La tasa de error aparente (TEA)} = \frac{0+9}{100} = 9\%$$

Es decir, usando la función discriminante lineal de Fisher, el 91% de las plantas de *Minthostachys* fueron bien clasificadas.

b.2. Resultados involucrados en el análisis de regresión lineal múltiple

Vectores de medias de plantas pubescentes y no pubescentes y el vector de medias combinado:

$$\vec{X}^{(1)} = \begin{bmatrix} 0.475 \\ 3.243 \\ 1.736 \end{bmatrix} \quad \text{y} \quad \vec{X}^{(2)} = \begin{bmatrix} 1.202 \\ 3.671 \\ 2.167 \end{bmatrix}$$

$$\vec{\overline{X}} = \begin{bmatrix} 1.2020 \\ 3.6714 \\ 2.1673 \end{bmatrix}$$

Matriz de datos centrados de las variables explicativas, consiste en quitarle a cada observación, el vector de medias combinado obtenido en el paso anterior, Luego, dichos datos, que se encuentran en el (anexo B), se usan como valores de las tres variables explicativas, en el ajuste del modelo de regresión lineal múltiple.

El vector de datos de la variable dependiente resulta,

$$Y_i^{(1)} = \frac{n_2}{n_1 + n_2} = \frac{51}{100} = 0.51 \qquad \text{y} \qquad Y_i^{(2)} = -\frac{n_1}{n_1 + n_2} = \frac{-49}{100} = -0.49$$

donde los 51 valores de la variable dependiente correspondiente al grupo de *Minthostachys* pubescentes, toman el valor 0.51 y los 49 valores de la variable dependiente correspondiente a las no pubescentes, toman el valor -0.49.

El vector de coeficientes de la recta de regresión lineal múltiple es:

$$\hat{\vec{\beta}} = \begin{bmatrix} -0.8456 \\ 0.1259 \\ 0.0416 \end{bmatrix}$$

Finalmente, se comprueba que $\hat{\vec{\beta}} = k\hat{\vec{\alpha}}$,

$$\hat{\beta} = \begin{bmatrix} -0.8456 \\ 0.1259 \\ 0.0416 \end{bmatrix} = k \quad \left[\hat{\alpha}\right] = \frac{1}{8.3983}\begin{bmatrix} -7.2491 \\ 1.0795 \\ 0.3566 \end{bmatrix}$$

donde $k = \frac{1}{8.3983}$.

c. Tercer conjunto de datos

El tercer conjunto de datos, tradicionalmente es usado en aplicaciones o cuando se abordan los temas de discriminación y clasificación (Fisher , 1936). En la presente aplicación se consideraron las 4 variables de las especies de iris Versicolor e iris Virginica respectivamente (Fisher, 1936**).**

El Tercer conjunto de datos está conformado por:

Grupo G_1: 50 de iris de tipo versicolor

Grupo G_2: 50 de iris de tipo de virginica

En ambos grupos se tomaron en cuenta las siguientes variables:

X_1 = Ancho de pétalo

X_2 = Largo de pétalo

X_3 = Ancho de sépalo

X_4 = Largo de sépalo.

c.1. Resultados involucrados en el análisis del análisis discriminante lineal

Vectores de medias y matrices de covarianzas de las especies iris versicolor e iris virgínica.

$$\vec{\bar{X}}^{(1)} = \begin{bmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.320 \end{bmatrix} \quad \text{y} \quad \vec{\bar{X}}^{(2)} = \begin{bmatrix} 6.5880 \\ 2.9740 \\ 5.5520 \\ 2.0260 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 0.266 & 0.085 & 0.183 & 0.056 \\ 0.085 & 0.098 & 0.083 & 0.041 \\ 0.183 & 0.083 & 0.221 & 0.073 \\ 0.056 & 0.041 & 0.073 & 0.039 \end{bmatrix} \quad \text{y} \quad S_2 = \begin{bmatrix} 0.404 & 0.094 & 0.303 & 0.049 \\ 0.049 & 0.104 & 0.071 & 0.048 \\ 0.0303 & 0.071 & 0.305 & 0.049 \\ 0.049 & 0.048 & 0.049 & 0.075 \end{bmatrix}$$

Matriz de covarianzas combinada:

$$S_u = \begin{bmatrix} 0.335 & 0.089 & 0.243 & 0.052 \\ 0.089 & 0.101 & 0.077 & 0.044 \\ 0.243 & 0.077 & 0.263 & 0.061 \\ 0.052 & 0.044 & 0.061 & 0.057 \end{bmatrix}$$

La prueba de hipótesis, para contrastar la hipótesis de igualdad de matrices de covarianzas, arrojó el siguiente valor para el estadístico M de Box(a través del software SPSS):

-2Log(M)	36.645
p-valor	0.000
Nivel de significación	0.05

Como el valor de la probabilidad asociada a la estadística M de Box , p-valor= 0.000, es menor que el nivel de significación prefijado, 0.05, se rechaza la hipótesis de que las matrices de covarianzas son iguales. Al nivel de significación 0.05, las matrices de varianzas y covarianzas son diferentes.

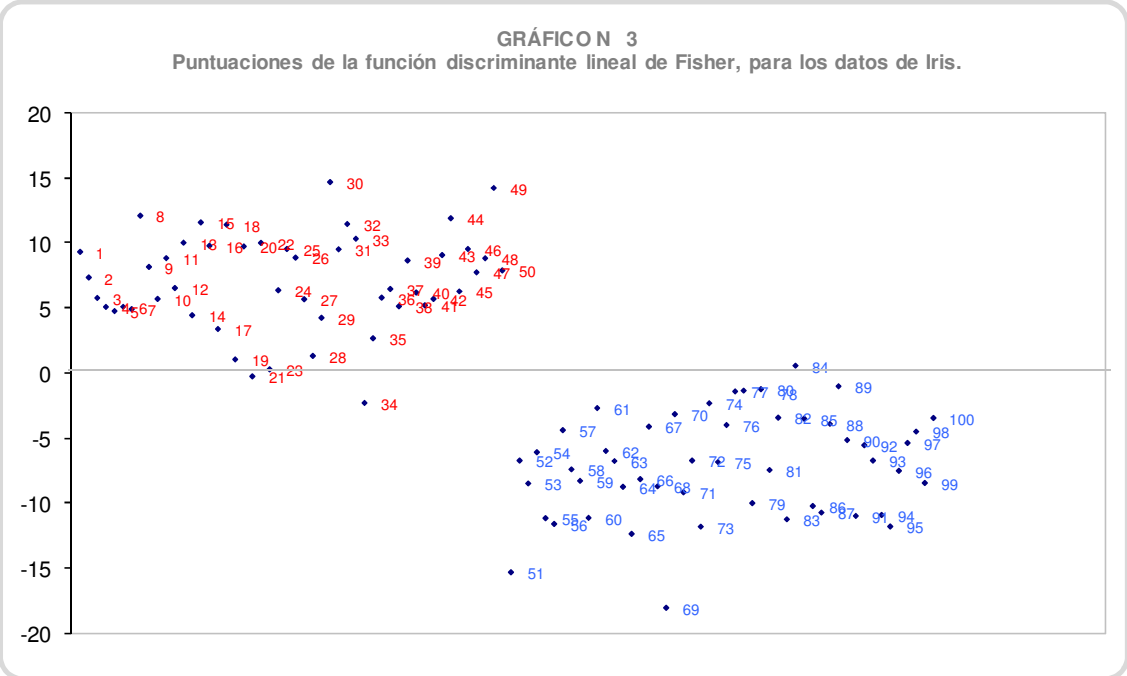
El vector de coeficientes de la función lineal discriminante de Fisher, (1.2.3), es:

$$\begin{bmatrix} \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 3.5561 \\ 5.5790 \\ -6.970 \\ -12.386 \end{bmatrix}$$

Las puntuaciones y la representación gráfica de la función discriminante de Fisher, se muestran en la TABLA Nº 3 el GRÀFICO Nº 3 respectivamente.

TABLA N° 3 Puntuaciones de la función discriminante lineal de Fisher, para los datos de iris									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
1	9,31	21	-0,25	41	5,20	61	-2,69	81	-7,45
2	7,33	22	9,99	42	5,69	62	-5,99	82	-3,42
3	5,76	23	0,28	43	9,05	63	-6,76	83	-11,24
4	5,07	24	6,35	44	11,89	64	-8,74	84	0,56
5	4,76	25	9,53	45	6,26	65	-12,36	85	-3,51
6	5,09	26	8,86	46	9,53	66	-8,15	86	-10,22
7	4,90	27	5,67	47	7,74	67	-4,12	87	-10,72
8	12,09	28	1,32	48	8,82	68	-8,70	88	-3,91
9	8,15	29	4,23	49	14,21	69	-18,03	89	-1,01
10	5,69	30	14,66	50	7,87	70	-3,16	90	-5,15
11	8,82	31	9,50	51	-15,31	71	-9,16	91	-10,98
12	6,53	32	11,44	52	-6,73	72	-6,73	92	-5,54
13	10,01	33	10,31	53	-8,49	73	-11,81	93	-6,73
14	4,43	34	-2,30	54	-6,08	74	-2,32	94	-10,91
15	11,56	35	2,66	55	-11,16	75	-6,84	95	-11,80
16	9,78	36	5,79	56	-11,59	76	-4,00	96	-7,51
17	3,37	37	6,45	57	-4,39	77	-1,42	97	-5,37
18	11,39	38	5,13	58	-7,40	78	-1,36	98	-4,50
19	1,04	39	8,63	59	-8,28	79	-10,00	99	-8,45
20	9,72	40	6,19	60	-11,13	80	-1,24	100	-3,46

Las observaciones mal clasificadas fueron: 21 y 34 de la especie Versicolor y 84 de la especie Virginica, lo que se evidencia mejor en el Gráfico Nº 3.



A continuación se presenta el resultado de la correspondiente clasificación.

Especie verdadera	Decisión estadística		
	Versicolor	Virgínica	Total
Versicolor	48	2	50
Virgínica	1	49	50

La $\text{tasa de error aparente}(\text{TEA}) = \frac{2+1}{100} x100 = 3\%$

Como se puede verificar, usando la función lineal discriminante lineal Fisher, el 97% de iris, perteneciente a ambas especies, fueron bien clasificadas. A pesar de no cumplir el supuesto de homoscedasticidad de las matrices de covarianzas, la clasificación es casi perfecta.

c.2. Resultados involucrados en el análisis de regresión lineal múltiple

Vectores de medias de cada grupo y el vector de medias combinado:

$$\vec{\overline{X}}^{(1)} = \begin{bmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.320 \end{bmatrix} \qquad \vec{\overline{X}}^{(2)} = \begin{bmatrix} 6.5880 \\ 2.9740 \\ 5.5520 \\ 2.0260 \end{bmatrix}$$

$$\vec{\overline{X}} = \begin{bmatrix} 6.262 \\ 2.872 \\ 4.906 \\ 1.673 \end{bmatrix}$$

Matriz de datos centrados de las variables explicativas:

Que consiste en quitarle a cada observación, el vector de medias combinado obtenido en el paso anterior, luego, dichos datos, que se encuentran en el (anexo B), se usan como valores de las (4) variables explicativas, en el ajuste del modelo de regresión lineal múltiple.

Vector de datos de la variable dependiente:

Usando la fórmula,

$$Y_i^{(1)} = \frac{n_2}{n_1 + n_2} = \frac{50}{100} \qquad \text{y} \qquad Y_i^{(2)} = -\frac{n_1}{n_1 + n_2} = \frac{-50}{100},$$

los 50 valores de la variable dependiente correspondiente al grupo 1, toman el valor 0.5 y los 50 valores de la variable dependiente correspondiente al grupo 2, toman el valor -0.5.

Vector de coeficientes de la recta de regresión lineal múltiple:

Al ajustar el modelo de regresión lineal múltiple con los datos centrados, como valores de las variables explicativas, a 0.5 y -0.5 para la variable dependiente, se estiman los parámetros de la recta de regresión lineal múltiple, resultando:

$$\hat{\beta} = \begin{bmatrix} 0.0196 \\ 0.0308 \\ -0.038 \\ -0.068 \end{bmatrix}$$

Finalmente, se comprueba que

$$\hat{\beta} = \begin{bmatrix} 0.0196 \\ 0.0308 \\ -0.038 \\ -0.068 \end{bmatrix} = k\left[\hat{\alpha}\right] = \frac{1}{14.1957}\begin{bmatrix} 0.35561 \\ 0.55790 \\ -0.6970 \\ -1.2386 \end{bmatrix}$$

donde $k = \frac{1}{14.1957}$.

Con lo cual se comprueba, también en este caso, la relación de proporcionalidad entre los coeficientes de la función lineal discriminante de Fisher y los de la función de regresión lineal múltiple.

Teniendo en cuenta los resultados obtenido con los tres conjunto de datos, efectivamente se comprueba que los coeficientes, de la función lineal discriminante de Fisher y de la función de regresión lineal múltiple, guardan una proporcionalidad.

CAPÍTULO II

OBSERVACIONES DISCORDANTES E INFLUYENTES

2.1. INTRODUCCIÓN

En todo análisis de datos, es de particular importancia la identificación de las observaciones denominadas discordantes, outliers, influyentes, que a veces se usan con una connotación similar sin serlo, que, dependiendo del caso, pueden producir grandes efectos en las estimaciones de los parámetros. Esta falta de estandarización en la terminología, es un inconveniente serio para lograr transmitir un mensaje claro al respecto y esto se ha acentuado aún más al existir una literatura amplia sobre este tema sin haber intentos de unificar los conceptos (Castaño, 1988). La misma traducción de outliers como "valor extremo" ya es una expresión propensa a crear confusión, pues un valor extremo no necesariamente es una observación discordante. Un artículo bastante completo sobre estas observaciones fue presentado por Beckman y Cook (1983), donde se hace un informe completo sobre las observaciones discordantes, aberrantes, contaminante, sorprendente, outliers**,** solo por nombrar algunos términos, pero la confusión aparece cuando en algunos estudios sobre este tópico, usan indiscriminadamente las expresiones anteriores.

En éste capítulo se presentan algunas definiciones que permiten precisar la diferencia entre observaciones discordantes y observaciones influyentes. También se presenta una breve revisión de trabajos que involucran observaciones discordantes o influyentes. Se muestra el efecto de una observación discordante en las estimaciones de los parámetros de posición central, de dispersión y de asociación (Peña, 2000).
Finalmente, se cuantifica los efectos que tiene una observación discordante en la estimación de los parámetros involucrados en el análisis discriminante lineal.

2.2. CONCEPTOS BÁSICOS

En el presente trabajo se tomará la definición sobre observaciones discordantes, basado en el comportamiento de la observación en relación a otras observaciones que fueron obtenidas en condiciones similares, en este contexto:

Definición 2.1 Una observación es discordante, cuando en opinión del investigador, se encuentra alejado de las demás observaciones que conforman el conjunto de datos motivo de análisis. También se le denomina aberrantes o discrepante*,* solo por mencionar algunos términos que le han asignados a lo largo de los años (Beckman y Cook,1983).

Definición 2.2 Una observación se denomina extrema, cuando en el contexto del análisis de regresión, su residuo es anormalmente grande (Anscombe,1963)[1].

Hay que tener clara las diferencias entre una observación extrema y una observación discordante, pues una observación denominada extrema no necesariamente es discordante.

Definición 2.3 Una observación influyente, es una observación discordante que al ser omitida en el análisis, da origen a cambios bruscos en las estimaciones de algunos y/o todos los parámetros involucrados en el estudio. Puede ser considerado como un caso especial de observación discordante.

*E*xistirán observaciones discordantes que no son influyentes, pues las estimaciones de los parámetros permanecen esencialmente inalterables cuando dicha observación es omitida (Beckman y Cook,1983).

Kotz y Jonson (1972)[2], textualmente dice:

"Las observaciones son consideradas como influyentes si su omisión da lugar a cambios sustanciales en aspectos importantes del análisis."

La presencia de las observaciones discordante se puede imputar a que la variable puede haber sido medida en una escala errada, el modelo propuesto no es el correcto, o errores en la toma de datos (Beckman y Cook ,1983).

[1] Referencia en Beckman, R.J. y Cook, R.D. (1983).
[2] Referencia en Enguix (2001)

Definición 2.4 En el contexto multivariante, una observación discordante, es aquella observación en el espacio multidimensional, que en opinión del investigador se encuentran alejada de las demás observaciones que conforman el conjunto de datos en análisis, esta observación es juzgada por el valor que toma no en una determinada variable, sino en el conjunto de ellas. Además, la identificación de observaciones discordantes multivariantes, es mucho más difícil que la identificación de observaciones discordantes univariantes y su presencia afecta especialmente las correlaciones entre las variables.

2.3. CUANTIFICACIÓN DEL EFECTO DE UNA OBSERVACIÓN DISCORDANTE EN EL ANÁLISIS MULTIVARIANTE

Para expresar matemáticamente el efecto de una observación discordante en las estimaciones del vector de medias y de la matriz de varianzas y covarianzas, vamos a suponer que:

$\vec{X}_1, \vec{X}_2, \vec{X}_3, ..., \vec{X}_a, ..., \vec{X}_n$ es una muestra aleatoria p dimensional de tamaño *n, donde*

$$\vec{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ . \\ . \\ . \\ X_{ip} \end{bmatrix} \quad i = 1,2,..., a, ..n$$

Dentro de la muestra existe una observación discordante que se denotada, $\vec{X}_a$,

$$\vec{X}_a = \begin{bmatrix} X_{a1} \\ X_{a2} \\ . \\ . \\ . \\ X_{ap} \end{bmatrix}$$

El vector de medias con toda la muestra, incluida la observación discordante es:

$$\overrightarrow{\overline{X}}_c = \frac{\sum_{i=1}^{n} \vec{X}_i}{n} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \cdot \\ \cdot \\ \overline{X}_p \end{bmatrix};$$

El vector de medias sin la observación discordante es:

$$\overrightarrow{\overline{X}} = \frac{\sum_{i=1}^{a-1} \vec{X}_i + \sum_{i=a+1}^{n} \vec{X}_i}{n-1}$$

La matriz de varianzas y covarianzas insesgada con la observación discordante

$$S_u^c = \frac{\sum_{i=1}^{n} \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right) \cdot \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right)^T}{n-1}$$

La matriz de covarianzas sin la observación discordante.

$$S_u = \frac{\sum_{i=1}^{a-1} \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right) \cdot \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right)^T + \sum_{i=a+1}^{n-1} \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right) \cdot \left(\vec{X}_i - \overrightarrow{\overline{X}}_c \right)^T}{n-2}$$

a) **Efecto de la observación discordante, en el vector de medias**

Por definición se tiene:

$$\overrightarrow{\overline{X}}_c = \frac{\sum_{i=1}^{n} \vec{X}_i}{n}$$

Aislando a la observación discordante de la sumatoria se llega a:

$$\overrightarrow{\overline{X}}_c = \frac{\sum_{i=1}^{n-1} \vec{X}_i + \vec{X}_a}{n}$$

$$= \frac{\frac{(n-1)\sum_{i=1}^{n-1} \vec{X}_i}{(n-1)} + \vec{X}_a}{n}$$

$$= \frac{(n-1)\overrightarrow{\overline{X}} + \vec{X}_a}{n}$$

$$\overrightarrow{\overline{X}}_c = \overrightarrow{\overline{X}} + \frac{\left(\vec{X}_a - \overrightarrow{\overline{X}}\right)}{n} \qquad (21)$$

Según el resultado anterior, la observación discordante no afectará al vector de medias, si $\vec{X}_a \approx \overrightarrow{\overline{X}}$ puesto que esta diferencia será nula o si el tamaño de muestra es suficientemente grande, en ambos casos el segundo término de (21) se eliminará y por ende se cumplirá que $\overrightarrow{\overline{X}}_c = \overrightarrow{\overline{X}}$.

$$\frac{\left(\vec{X}_a - \overrightarrow{\overline{X}}\right)}{n} \to 0$$

b) Efecto de la observación discordante, la matriz de varianzas y covarianzas

Otro de los parámetros que se queda afectado con la presencia de una observación discordante en el conjunto de datos, es la matriz de varianzas y covarianzas. A continuación se cuantifica el efecto de la observación discordante en la estimación de dicho parámetro.

En efecto, por definición se tiene:

$$S_u^c = \frac{\sum_{i=1}^{n}\left(\vec{X}_i - \overrightarrow{\overline{X}}_c\right).\left(\vec{X}_i - \overrightarrow{\overline{X}}_c\right)^T}{n-1}$$

Aislando de la sumatoria, el término que contiene la diferencia de la observación discordante y el vector de medias calculada con dicha observación.

$$S_u^c = \frac{\sum_{i=1}^{n-1}\left(\vec{X}_i - \overrightarrow{\overline{X}}_c\right).\left(\vec{X}_i - \overrightarrow{\overline{X}}_c\right)^T + \left(\vec{X}\,a - \overrightarrow{\overline{X}}_c\right).\left(\vec{X}\,a - \overrightarrow{\overline{X}}_c\right)^T}{n-1}$$

Reemplazando el valor del vector de medias afectada por la observación discordante, dada en (21).

$$S_u^c = \frac{\sum_{i=1}^{n-1}\left(\vec{X}_i - \overrightarrow{\overline{X}} - \frac{\vec{X}_a - \overrightarrow{\overline{X}}}{n}\right).\left(\vec{X}_i - \overrightarrow{\overline{X}} - \frac{\vec{X}_a - \overrightarrow{\overline{X}}}{n}\right)^T + \left(\vec{X}\,a - \overrightarrow{\overline{X}} - \frac{\vec{X}_a - \overrightarrow{\overline{X}}}{n}\right).\left(\vec{X}\,a - \overrightarrow{\overline{X}} - \frac{\vec{X}_a - \overrightarrow{\overline{X}}}{n}\right)^T}{n-1}$$

$$S_u^c = \frac{\sum\left[\left(\vec{X}_i-\vec{\bar{X}}\right).\left(\vec{X}_i-\vec{\bar{X}}\right)^T-\left(\vec{X}_i-\vec{\bar{X}}\right)\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right)^T-\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right).\left(\vec{X}_i-\vec{\bar{X}}\right)^T+.\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right).\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right)^T\right]}{n-1}$$

$$+\frac{\left(\frac{n.\vec{X}_a-n.\vec{\bar{X}}-\vec{X}_a+\vec{\bar{X}}}{n}\right).\left(\frac{n.\vec{X}_a-n.\vec{\bar{X}}-\vec{X}_a+\vec{\bar{X}}}{n}\right)^T}{n-1}$$

Desarrollando la sumatoria

$$S_u^c = \frac{\sum_{i=1}^{n-1}\left(\vec{X}_i-\vec{\bar{X}}\right).\left(\vec{X}_i-\vec{\bar{X}}\right)^T-\sum_{i=1}^{n-1}\left[\left(\vec{X}_i-\vec{\bar{X}}\right)\right]\left[\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right)^T\right]-\left[\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right)\right]\sum_{i=1}^{n-1}\left[\left(\vec{X}_i-\vec{\bar{X}}\right)^T\right]}{n-1}$$

$$+\frac{(n-1).\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right).\left(\frac{\vec{X}_a-\vec{\bar{X}}}{n}\right)^T+\left(\frac{(n-1).\left(\vec{X}_a-\vec{\bar{X}}\right)}{n}\right).\left(\frac{(n-1).\left(\vec{X}_a-\vec{\bar{X}}\right)}{n}\right)^T}{n-1}$$

Dando forma a la expresión equivalente de S, además de aplicar algebra elemental

$$S_u^c = \frac{\frac{(n-2).\left[\sum_{i=1}^{n-1}\left(\vec{X}_i-\vec{\bar{X}}\right).\left(\vec{X}_i-\vec{\bar{X}}\right)^T\right]}{(n-2)}+\frac{(n-1)}{n^2}.\left(\vec{X}_a-\vec{\bar{X}}\right).\left(\vec{X}_a-\vec{\bar{X}}\right)^T+\frac{(n-1)^2}{n^2}\left(\vec{X}_a-\vec{\bar{X}}\right)\left(\vec{X}_a-\vec{\bar{X}}\right)^T}{n-1}$$

Reemplazando el valor de S, se obtiene

$$S_u^c = \frac{(n-2).S+\frac{(n-1)}{n^2}+\frac{(n-1)}{n}\left(\vec{X}_a-\vec{\bar{X}}\right)\left(\vec{X}_a-\vec{\bar{X}}\right)^T}{n-1}$$

Luego de algunas simplificaciones se llega a:

$$S_u^c = \frac{(n-2)}{(n-1)}S_u+\frac{\left(\vec{X}_a-\vec{\bar{X}}\right)\left(\vec{X}_a-\vec{\bar{X}}\right)^T}{n} \qquad (22)$$

que causa una observación discordante en las estimaciones de los parámetros:

Para el caso de la matriz de varianzas y covarianzas, el efecto de la observación discordante será nulo si $\vec{X}_a \approx \vec{\bar{X}}$ debido a que la diferencia será nula y además para un tamaño de muestra suficientemente grande, el segundo término de (22) es:

$$\frac{\left(\vec{X}_a - \vec{\overline{X}}\right)\left(\vec{X}_a - \vec{\overline{X}}\right)^T}{n} \to 0 \qquad \text{y además } \frac{(n-2)}{(n-1)} \to 1$$

Con lo que $S_u^c = S_u$, que puede ser interpretado que en las condiciones mencionadas el efecto de la observación discordante será nula.

2.4. CUANTIFICACIÓN DEL EFECTO DE UNA OBSERVACIÓN DISCORDANTE EN EL ANÁLISIS DISCRIMINANTE LINEAL EN DOS GRUPOS

Llevando en cuenta, los resultados obtenidos en la sección anterior, se presentará el efecto que tiene una observación discordante en las estimaciones de los parámetros involucrados en el análisis discriminante lineal en dos grupos, como son:
La distancia de Mahalanobis, los promedios discriminantes, los coeficientes de la función discriminante de Fisher, entre otros.

Para cuantificar el efecto de la observación discordante en las estimaciones de los parámetros de interés, en el análisis discriminante lineal en dos grupos, se tomará en cuenta lo siguiente:

Supongamos que: $\vec{X}_1^{(k)}, \ldots\ldots, \vec{X}_{n_k}^{(k)}$; $k = 1,2$ y $n_1 + n_2 = n$ muestras aleatorias de cada uno de los grupos o poblaciones,

Una observación discordante está dentro del primer grupo, denotado como $\vec{X}_a^{(1)^T}$, donde

$$\vec{X}_a^{(1)^T} = \begin{bmatrix} X_{a1}^{(1)} \\ X_{a2}^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ X_{ap}^{(1)} \end{bmatrix},$$

$$\overrightarrow{\overline{X}}_c^{(1)} = \frac{\sum_{i=1}^{n_1} \overrightarrow{X}_i^{(1)}}{n_1} = \begin{bmatrix} \overline{X}_1^{(1)} \\ \overline{X}_2^{(1)} \\ . \\ . \\ \overline{X}_p^{(1)} \end{bmatrix}$$, vector de medias del primer grupo con la observación discordante,

$$\overrightarrow{\overline{X}}^{(1)} = \frac{\sum_{i=1}^{n_1-1} \overrightarrow{X}_i^{(1)}}{n_1 - 1}$$ Vector de medias del primer grupo sin la observación discordante,

$$S_1^c = \frac{\sum_{i=1}^{n_1} \left(\overrightarrow{X}_i^{(1)} - \overrightarrow{\overline{X}}_c^{(1)} \right) . \left(\overrightarrow{X}_i^{(1)} - \overrightarrow{\overline{X}}_c^{(1)} \right)^T}{n_1 - 1}$$ Matriz de varianzas y covarianzas del primer grupo con la observación discordante,

$$S_1^c = \frac{\sum_{i=1}^{n_1-1} \left(\overrightarrow{X}_i^{(1)} - \overrightarrow{\overline{X}}^{(1)} \right) . \left(\overrightarrow{X}_i^{(1)} - \overrightarrow{\overline{X}}^{(1)} \right)^T}{n_1 - 2}$$ Matriz de varianzas y covarianzas del primer grupo sin la observación discordante,

Llevando en cuenta, los resultados mostrados en (21) y en (22) se tiene:

$$\overrightarrow{\overline{X}}_c^{(1)} = \overrightarrow{\overline{X}}^{(1)} + \frac{\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)} \right)}{n_1} \qquad (23)$$

$$S_1^c = \frac{(n_1 - 2)}{(n_1 - 1)} S_1 + \frac{\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)} \right) . \left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)} \right)^T}{n_1} \qquad (24)$$

a. Efecto de la observación discordante en la matriz de covarianzas combinada

Reemplazando (24) en la matriz de varianzas y covarianza combinada de obtiene:

$$S_u^c = \frac{(n_1-1).\left[\frac{(n_1-2)}{(n_1-1)}S_1^c + \frac{\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right).\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right)^T}{n_1}\right] + (n_2-1)S_2}{n_1+n_2-2} \qquad (25)$$

que es la matriz de varianzas y covarianzas combinada, afectada por la observación discordante.

El efecto causado será nulo si ocurre lo siguiente:

$$\frac{\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right).\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right)^T}{n_1} \to 0 \qquad \text{y} \qquad \frac{(n_1-2)}{(n_1-1)} \to 1$$

b. Efecto de la observación discordante en la distancia de Mahalanobis

Otro de los parámetros que se ve afectado por la presencia de una observación discordante es la distancia de Mahalanobis, para cuantificar el efecto, se reemplaza (23) y (25), en la definición de dicha distancia, así se obtiene:

$$D_c^2 = \left(\overrightarrow{\overline{X}}^{(1)} + \frac{\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}}{n_1} - \overrightarrow{\overline{X}}^{(2)}\right)^T .\left(S_u^c\right)^{-1}.\left(\overrightarrow{\overline{X}}^{(1)} + \frac{\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}}{n_1} - \overrightarrow{\overline{X}}^{(2)}\right)$$

El efecto de la observación será nulo si al igual que en (21)

$$\frac{\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right).\left(\overrightarrow{X}_a^{(1)} - \overrightarrow{\overline{X}}^{(1)}\right)^T}{n_1} \to 0 \qquad \text{y} \qquad \frac{(n_1-2)}{(n_1-1)} \to 1$$

Hay que mencionar también, que el efecto que ocasiona la observación discordante en la distancia de Mahalanobis, se verá reflejado en las probabilidades de mala clasificación.

c. Efecto de la observación discordante en los coeficientes de la función discriminante lineal de Fisher

El efecto de la observación discordante en los coeficientes de la función discriminante lineal de Fisher, puede cuantificarse reemplazando en la definición de dicho coeficiente, los resultados mostrados en (23) y en (24) si se obtiene:

$$\hat{\vec{\alpha}}_c = \left(S_u^c\right)^{-1}.\left(\vec{\overline{X}}^{(1)} + \frac{\vec{X}_a^{(1)} - \vec{\overline{X}}^{(1)}}{n_1} - \vec{\overline{X}}^{(2)}\right)$$

El efecto, de la observación discordante, será nulo al igual que los anteriores si

$\frac{\left(\vec{X}_a^{(1)} - \vec{\overline{X}}^{(1)}\right).\left(\vec{X}_a^{(1)} - \vec{\overline{X}}^{(1)}\right)^T}{n_1} \to 0$ y $\frac{(n_1-2)}{(n_1-1)} \to 1$ y esto será posible si se tiene un tamaño de muestra considerablemente grande, y además sí $\vec{\overline{X}}_a^{(1)} = \vec{\overline{X}}^{(1)}$.

Los diferentes resultados encontrados en este capítulo, en referencia al efecto que puede ocasionar una observación discordante en las estimaciones de los distintos parámetros de interés, nos permite sacar las siguientes conclusiones:

- Mientras se tenga un tamaño de muestra suficientemente grande y además $\vec{X}_a \approx \vec{\overline{X}}$, el efecto de una observación discordante será nula, o no será de cuidado.
- En el caso que no se cuenta con un tamaño de muestra suficientemente o que $\vec{X}_a$ difiera considerablemente de $\vec{\overline{X}}$, el efecto de dicha observación será tremendamente perjudicial o tendrá bastante influencia en las estimaciones de los distintos parámetros de interés, es necesario entonces desarrollar métodos y/o técnicas que permitan identificar a estas observaciones, precisamente en el siguiente capítulo se desarrollan algunas de estas medidas.

CAPÍTULO III

MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL EN DOS GRUPOS

3.1. INTRODUCCIÓN

En el contexto de los modelos de regresión, se han desarrollado y propuesto medidas para identificar las denominadas observaciones influyentes, así como para cuantificar el cambio o efecto que ellas originan en la estimación de los parámetros de los modelos, muchas de las cuales han sido incorporadas en algunos software estadísticos y como referencias recientes importantes pueden citarse (Belsley et al.,2004; Atkinson et. al., 2004 & Maronna et al.,2006).

En el marco de los modelos multivariantes, (Gnanadesikan y Kettenring1972)[3]; señalan que las observaciones discordantes multidimensionales son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de ellas, además son mucho más difíciles de identificar que los outliers unidimensionales, dado que no pueden considerarse "valores extremos", como sucede cuando se tiene una única variable, su presencia tiene efectos todavía más perjudiciales que en el caso unidimensional, porque distorsionan no sólo los valores de la medida de posición (media) o de dispersión (varianza), sino, muy especialmente, las correlaciones entre las variables, que es precisamente la idea principal en la que se basan estos tipo de estudios (Morillas y Diaz, 2007).

En cuanto al análisis discriminante, a pesar de que la estimación de parámetros es un tema ampliamente estudiado y difundido (Mardia,1979; Johnson, 2000; Hair et al.,1999 & Manly, 2005, entre otros) pocas son las propuestas sostenidas respecto al estudio sobre la identificación de las observaciones influyentes.

Uno de los primeros trabajos referidos a este tema fue presentado por Campbell (1978), en el que propone medidas de influencia para diversos parámetros de interés, basándose en la función de influencia propuesta por Hampel (1974); años más tarde Cook y Weisberg (1982) propusieron una medida referente a la probabilidad de mala clasificación, basándose en la aproximación de la función de influencia realizada por Devlín et. al. (1976), que luego Fung

[3] Referencia en Campbell (1978)

(1992), propondría una modificación de esta medida. Radhakrishnan [4] (1985), también propuso funciones de influencia para detectar observaciones influyentes correspondientes a funciones discriminantes para múltiples grupos y finalmente Fung (1995), apoyándose en la relación que existe entre los coeficientes de la función discriminante lineal de Fisher y los coeficientes del modelo de regresión lineal múltiple, realizó un paralelo de cómo se construye las medidas de influencia en el análisis de regresión lineal, para proponer de manera similar dos medidas de influencia en el análisis discriminante lineal.

En este contexto, lo que se pretende mostrar en esta sección, es el desarrollo teórico y metodológico de algunas funciones de influencia.

Para ello, este capítulo se ha dividido en tres secciones. En la primera se hace la presentación del análisis de influencia, enfoque que nos servirá para la propuesta de las medidas de influencia. Seguidamente se hace una breve descripción de la función de influencia propuesto por Hampel (1974), y algunas alternativas a ella propuesta por algunos investigadores; y finalmente en la tercera sección se presentan las funciones de influencia para los principales parámetros del análisis discriminante lineal.

3.2. ANÁLISIS DE INFLUENCIA

Uno de los problemas que muchas veces enfrenta un estadístico, al realizar el análisis de un conjunto de datos, es la presencia de una o un conjunto de observaciones influyentes. Frente a este problema, se han propuesto un conjunto de técnicas y/o métodos para detectarlas y en la literatura, generalmente se presentan bajo la denominación de análisis de influencia.

En gran medida el análisis de influencia, ha sido desarrollado a partir de la función de influencia propuesta por Hampel (1974), que introdujo dicha función con el objetivo de estudiar la robustez de los estimadores.

La idea básica del análisis de influencia consiste en introducir pequeñas perturbaciones en la formulación del problema y evaluar los resultados obtenidos por efecto de la perturbación. Es decir, comparar las estimaciones obtenidas con y sin la perturbación. En diversos estudios realizados sobre este tema, el tipo de perturbación más usado es el de la omisión de observaciones.

[4] Referencia en Fung (1995)

Muños et. al. (2001), han propuesto el siguiente esquema, que engloba una idea clara de lo que es el análisis de influencia.

- **D** un conjunto de datos,
- **M** un modelo postulado a priori,
- **R(D;M)** un resultado**,** seleccionado de una síntesis de los datos y el modelo,
- **w** un vector de perturbaciones, perteneciente a un conjunto **Ω** de perturbaciones relevantes,
- **M(w)** el modelo perturbado, de forma que

$$\exists w_0 \in \Omega / M \approx M(w_0)$$

- **R(D;M(w)):** el resultado obtenido con la perturbación,

Entonces, el análisis de influencia consiste en comparar los resultados de

R(D,M(w)) y ***R(D,M).***

3.3. FUNCIÓN DE INFLUENCIA

El concepto de función de influencia, fue introducido por Hampel (1974), formalmente es definida como:

Se considera un parámetro general $\theta = T(F_1,...,F_k,...,F_g)$ expresado como funcional de la función de distribución F_g, donde $g = 1,2,...,k$. La función de distribución perturbada puede ser escrito como:

$$\tilde{\tilde{F}}_k = (1-\varepsilon)F_k + \varepsilon\delta_{\underset{x}{\rightarrow}}$$

donde, $\delta_{\underset{x}{\rightarrow}}$ es a la función de distribución de la variable aleatoria degenerada o perturbada en el punto $\vec{x} = (x_1, x_2, ..., x_p)$.

Sea $\tilde{\tilde{\theta}}_k = T\left(\tilde{\tilde{F}}_1,...\tilde{\tilde{F}}_k,...,\tilde{\tilde{F}}_g\right)$ el parámetro de evaluada en la función de distribución de perturbada y además $\forall\varepsilon \in (0,1)$, la función de influencia en el punto $\vec{x} = (x_1, x_2, ..., x_p)$ está definida como:

$$I_k\left(\vec{x};\theta\right)=\lim_{\varepsilon\to 0}\frac{\tilde{\tilde{\theta}}-\theta}{\varepsilon} \qquad (26)$$

Para todo $\vec{x}$ del espacio muestral en que existe el limite.

El subíndice "k" no se conserva en el resto del documento, ya que sólo la distribución de la primera población es perturbada.

3.3.1. Estimación de la función de influencia

En la práctica, el objetivo del análisis de influencia es evaluar el efecto que tiene una observación sobre un estadístico determinado, entonces es necesario las versiones muéstrales de la función de influencia, según la literatura revisada las más utilizadas son las versiones propuestas por Mallows (1976)[5] que se presenta en forma resumida a continuación.

Sea $\vec{X}_1\,\vec{X}_2,\vec{X}_3,...,\vec{X}_n$ una muestra aleatoria de un vector aleatorio $\vec{X}$ p-dimensional, se denota como $\hat{F}_n$ a la función de distribución muestral basada en la realización de la misma y a $\hat{F}_{n-1}^{(i)}$ como la función de distribución muestral omitiendo la i-ésima observación de la muestra, entonces se define la función de influencia muestral de la siguiente manera.

Definición 3.1 Sean $\vec{X}_1\,\vec{X}_2,\vec{X}_3,...,\vec{X}_n$ una muestra aleatoria procedente de una variable aleatoria $\vec{X}$ p-dimensional, y sea $x_1,x_2,x_3,...,x_n$ la realización muestral de la misma. La función de influencia muestral se define como:

$$FIE\left(\vec{x}_i;T\left(\hat{F}_n\right)\right)=\hat{I}\left(\vec{x};\hat{F}_n\right)$$

Función de influencia muestral con omisión como:

$$FIE\left(\vec{x}_i;T\left(\hat{F}_{n-1}^{(i)}\right)\right)=\hat{I}\left(\vec{x};\hat{F}_{n-1}^{(i)}\right)$$

[5] Referencia en Enguix (2001)

3.3.2. Aproximación de la función de influencia

La siguiente versión muestral fue propuesta en primer lugar por Mallows (1975), aunque su nombre se deba a Devlín et al. (1975)[6]. Se basa en el hecho de que, para un tamaño muestral suficientemente grande, la función de influencia muestral se puede considerar como una aproximación de la función de influencia eliminando el límite y tomando como función de distribución, la función de distribución muestral $\hat{F}_n$ y $\varepsilon = -\frac{1}{n-1}$, de este modo la perturbación de $\hat{F}_n$ es la función de distribución de la muestra de tamaño $(n-1)$ obtenida al eliminar la i-ésima observación, $\overset{\approx}{\hat{F}}_{n-1}^{(i)}$

$$\left[1-\left(-\frac{1}{n-1}\right)\right].\hat{F}_n+\left(-\frac{1}{n-1}\right).\delta_{\underset{x_i}{\rightarrow}} = \frac{n.\hat{F}_n-.\delta_{\underset{x_i}{\rightarrow}}}{n-1} = \overset{\approx}{\hat{F}}_{n-1}^{(i)} \qquad (27)$$

3.3.3. Función de influencia general

Esta función de influencia, es más general al planteamiento realizado previamente, al definirse un estadístico cualquiera, sin necesidad de un funcional que lo determine, (Enguix, 2001), así la función de influencia general está definida como:

Sea $\vec{X}_1\,\vec{X}_2,\vec{X}_3,...,\vec{X}_n$ una muestra aleatoria procedente de una variable aleatoria $\vec{X}$ p-dimensional, y sea $x_1,x_2,x_3,...,x_n$ la realización muestral de la misma, se define la función la función de influencia muestral de un estadístico $T=T\left(\vec{X}_1\,\vec{X}_2,\vec{X}_3,...,\vec{X}_n\right)$ como:

$$FIM\left(\vec{x}_i;T\right)=(n-1).\left[T-T^{(i)}\right] \qquad (28)$$

Que básicamente puede interpretarse como una comparación de los estadísticos calculados sobre la muestra completa y la muestra sin la i-ésima observación.

3.4. MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL EN DOS GRUPOS

Uno de los primeros en realizar estudios sobre influencia, en el análisis discriminante lineal basándose en la función de influencia introducida por Hampel (1974), fue Campbell (1978), quien propuso medidas de influencia para diversos parámetros de interés, como la distancia de Mahalanobis, promedios discriminantes, vector de coeficientes, que fueron definidas anteriormente, además sugiere que antes de definir las funciones de influencia, debe considerarse el efecto de la perturbación en $\vec{\mu}^{(k)}$ y Σ^{-1} donde:

$$\Sigma = w_1.\Sigma_{F_1} + w_2.\Sigma_{F_2}$$

Para $w_1 + w_2 = 1$ y $w_k > 0$, además

$$\Sigma_{F_k} = \int \left(\vec{x} - \vec{\mu}^{(k)}\right)\left(\vec{x} - \vec{\mu}^{(k)}\right)^T d_{F_k}$$

$$\vec{\mu}^{(k)} = \int \vec{x} d_{F_k}$$

En la derivación de las funciones de influencia, se asume que $\Sigma_1 = \Sigma_2$

En primer lugar se mostrará el efecto de la perturbación en $\vec{\mu}^{(k)}$ y Σ^{-1}, solo por efecto de hacer entendible la deducción de la medida, se asumirá la perturbación se hará solo en el grupo G_1

- Para $\vec{\mu}^{(1)}$

$$\vec{\tilde{\mu}}^{(1)} \to (1-\varepsilon)\vec{\mu}^{(1)} + \varepsilon.\vec{x}$$

$$\vec{\mu}^{(1)} + \varepsilon.\left(\vec{x} - \vec{\mu}^{(1)}\right)$$

Sea $\vec{x} - \vec{\mu}^{(1)} = \vec{Z}$, entonces

$$\vec{\tilde{\mu}}^{(1)} \to \vec{\mu}^{(1)} + \varepsilon.\vec{z} \qquad (29)$$

- Para $\left(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}\right) = \vec{\delta}$

Siguiendo el mismo argumento que se utilizó para hallar $\overset{\approx}{\vec{\mu}}^{(1)}$ se tiene lo siguiente

$$\overset{\approx}{\vec{\delta}} \to \vec{\delta} + \varepsilon . \vec{Z} \qquad (30)$$

- Para Σ_{1_1}

$$\overset{\approx}{\Sigma}_1 \to (1-\varepsilon)\Sigma_1 + \varepsilon \left(\vec{x} - \vec{\mu}^{(1)} \right)\left(\vec{x} - \vec{\mu}^{(1)} \right)^T$$

$$(1-\varepsilon)\Sigma_1 + \varepsilon . \vec{Z} . \vec{Z}^T$$

- Para Σ

$$\overset{\approx}{\Sigma} \to (1 - w_1 \varepsilon)\Sigma_1 + \varepsilon w_1 \vec{Z}\, \vec{Z}^T$$

- Para Σ^{-1} como: (véase parte de este resultado en el apéndice A)

$$\overset{\approx}{\Sigma}^{-1} \to (1-\varepsilon w_1)\Sigma^{-1} - \varepsilon w_1 \Sigma^{-1} \vec{Z} . \vec{Z}^T . \Sigma^{-1} \qquad (31)$$

3.4.1. Medida de influencia para la distancia de Mahalanobis

Llevando en cuenta los resultados obtenidos en (30) y (31) la perturbación en Δ^2 es expresado como:

$$\overset{\approx}{\Delta}^2 \to \left(\vec{\delta} + \varepsilon . \vec{Z} \right)^T \left[(1 + \varepsilon w_1)\Sigma^{-1} - \varepsilon w_1 \Sigma^{-1} \vec{Z} . \vec{Z}^T . \Sigma^{-1} \right] \left(\vec{\delta} + \varepsilon \vec{Z} \right)$$

Aplicando propiedades de vectores y matrices siempre que estén bien definidas la multiplicación de las mismas, se tiene:

$$\overset{\approx}{\Delta}^2 \to \left(\vec{\delta} + \varepsilon . \vec{Z} \right)^T [(1+\varepsilon w_1).\Sigma^{-1}] \left(\vec{\delta} + \varepsilon . \vec{Z} \right) - \left(\vec{\delta} + \varepsilon . \vec{Z} \right)^T \left[\varepsilon w_1 . \Sigma^{-1} \vec{Z} . \vec{Z}^T . \Sigma^{-1} \right] . \left(\vec{\delta} + \varepsilon . \vec{Z} \right)$$

$$\overset{\approx}{\Delta}^2 \to (1+\varepsilon w_1) \left[\left(\vec{\delta} + \varepsilon . \vec{Z} \right)^T \Sigma^{-1} \left(\vec{\delta} + \varepsilon . \vec{Z} \right) \right] - \varepsilon w_1 . \left[\left(\vec{\delta} + \varepsilon . \vec{Z} \right)^T \Sigma^{-1} \vec{Z} . \vec{Z}^T . \Sigma^{-1} . \left(\vec{\delta} + \varepsilon . \vec{Z} \right) \right]$$

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1)\left[\vec{\delta}^T\,\Sigma^{-1}\vec{\delta}+\vec{\delta}^T\,\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)+\left(\varepsilon.\vec{Z}\right)^T\Sigma^{-1}\vec{\delta}+\left(\varepsilon.\vec{Z}\right)^T\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)\right]-$$

$$\varepsilon w_1.\left[\vec{\delta}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T.\Sigma^{-1}\vec{\delta}+\vec{\delta}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T.\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)+\left(\varepsilon.\vec{Z}\right)^T\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\vec{\delta}+\left(\varepsilon.\vec{Z}\right)^T\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)\right]$$

Reemplazando las expresiones $\Delta^2=\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)^T\Sigma^{-1}\left(\vec{\mu}^{(1)}-\vec{\mu}^{(2)}\right)=\vec{\delta}^T\,\Sigma^{-1}\vec{\delta}$

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1)\left[\Delta^2+\varepsilon\left(\vec{\delta}^T\,\Sigma^{-1}\vec{Z}\right)+\varepsilon\left(\vec{\delta}\Sigma^{-1}\vec{Z}\right)^T+\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]-$$

$$\varepsilon w_1.\left[\vec{\delta}^T\,\Sigma^{-1}\vec{Z}\left(\vec{\delta}^T\,\Sigma^{-1}\vec{Z}\right)^T+\left(\vec{\delta}^T\,\Sigma^{-1}\vec{Z}\right).\vec{Z}^T.\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)+\varepsilon.\vec{Z}\Sigma^{-1}\vec{Z}.\left(\vec{\delta}^T\,\Sigma^{-1}\vec{Z}\right)^T+\varepsilon\vec{Z}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]$$

También haciendo $\psi=\vec{\delta}^T\,\Sigma^{-1}\vec{Z}$

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1)\left[\Delta^2+\varepsilon\psi+\varepsilon(\psi)^T+\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]-$$

$$\varepsilon w_1.\left[\psi(\psi)^T+(\psi).\vec{Z}^T.\Sigma^{-1}\left(\varepsilon.\vec{Z}\right)+\varepsilon.\vec{Z}\Sigma^{-1}\vec{Z}.(\psi)^T+\varepsilon\vec{Z}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]$$

Como ψ está definida en R $\psi=\psi^T$, entonces:

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1)\left[\Delta^2+2\varepsilon\psi+\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]-\varepsilon w_1.\left[\psi^2+2\varepsilon\psi\vec{Z}^T\,\Sigma^{-1}\vec{Z}+\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\vec{Z}\right]$$

Aplicando propiedades de algebra elemental se obtiene:

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1)\Delta^2+2\varepsilon\psi+2\varepsilon^2\psi+\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}+\varepsilon^3 w_1\vec{Z}^T\,\Sigma^{-1}\vec{Z}-.\varepsilon w_1\psi^2$$
$$-2\varepsilon^2 w_1\psi\vec{Z}^T\,\Sigma^{-1}\vec{Z}-\varepsilon^2\vec{Z}^T\,\Sigma^{-1}\vec{Z}.\vec{Z}^T\,\Sigma^{-1}\vec{Z}$$

Solo reteniendo los términos de orden ε y dejando de lado los demás términos se tiene:

$$\tilde{\Delta}^2 \to (1+\varepsilon w_1).\Delta^2+2.\varepsilon\psi-w_1.\psi^2 \qquad (32)$$

Evaluando (32) en (26)

$$I\left(\vec{x};\Delta^2\right)=\lim_{\varepsilon\to 0^+}\frac{(1+\varepsilon w_1).\Delta^2+2.\varepsilon.\psi-w_1.\psi^2-\Delta^2}{\varepsilon}$$

$$I\left(\vec{x};\Delta^2\right)=\lim_{\varepsilon\to 0^+}\left(\frac{1}{\varepsilon}\Delta^2+w_1\Delta^2+2\psi-w_1\psi^2-\frac{1}{\varepsilon}\Delta^2\right)$$

Finalmente la medida de influencia para la distancia de Mahalanobis es:

$$I\left(\vec{x};\Delta^2\right) = w_1\Delta^2 + 2\psi - w_1\psi^2 \qquad (33)$$

La medida de influencia para Δ^2, mostrada en (33), tiene la dificultad cuando el resultado de $w_1\psi^2$ es superior a $\left(w_1\Delta^2 + 2\psi\right)$, en estos casos los valores que toma $I\left(\vec{x};\Delta^2\right)$ será negativo caso contrario los valores de es $I\left(\vec{x};\Delta^2\right)$ serán positivos, en estas circunstancias no es tan sencillo etiquetar a una observación como influyente, por tal motivo se propone la siguiente modificación:

$$I_M\left(\vec{x};\Delta^2\right) = \max_x\left[I\left(\vec{x};\Delta^2\right)\right] - I\left(\vec{x};\Delta^2\right) \qquad (34)$$

Donde el máximo de $I_M\left(\vec{x};\Delta^2\right)$ es $I_{\max}\left(\vec{x};\Delta^2\right) = w_1.\Delta^2 + w_1^{-1}$, (véase este resultado desarrollado en el apéndice A), con lo cual se tiene lo siguiente:

$$I_M\left(\vec{x};\Delta^2\right) = \left(w_1.\Delta^2 + w_1^{-1}\right) - \left(w_1\Delta^2 + 2\psi - w_1\psi^2\right)$$

Que es equivalente a:

$$I_M\left(\vec{x};\Delta^2\right) = w_1\left[\psi - w_1^{-1}\right]^2 \qquad (35)$$

Estimación de la medida de influencia para la distancia de Mahalanobis

Sea $\vec{X}_1, \vec{X}_2, \vec{X}_3, ..., \vec{X}_{n_k}$; $k = 1,2$ y $n_1 + n_2 = n$, muestras aleatorias de la variable aleatoria $\vec{X}$ p-dimensional, como ya de mencionó anteriormente en el análisis discriminante lineal, generalmente se considera como estimador de la media poblacional $\vec{\mu}^{(k)}$ a la media muestral $\vec{\overline{X}}^{(k)}$ y como estimador de la matriz de varianzas y covarianzas a S_u, según esto la estimación de la función de influencia para Δ^2 es dada como:

$$\hat{I}_M\left(\vec{x};\Delta^2\right) = w_1\left[\hat{\psi} - w_1^{-1}\right]^2 \qquad (36)$$

donde: $\hat{\alpha} = S_u^{-1}.\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}^{(2)}\right)$ y $\hat{\psi} = \hat{\alpha}^T.\left(\vec{x} - \vec{\overline{X}}^{(1)}\right)$

Observando la estructura de la medida presentada en (36), podemos concluir, que esta medida depende básicamente de la estadística $\hat{\psi}$, que será explicado con detalle más adelante.

3.4.2. Medida de influencia para la probabilidad de mala clasificación

Como ya se mencionó, una de las cruciales consideraciones en el análisis discriminante, es la probabilidad de mala clasificación, bajo la regla discriminante lineal de la población propuesta por Fisher, la mencionada probabilidad de mala clasificación está definida como: $P(i/j;R) = \phi\left(-\frac{1}{2}\Delta\right)$

La función de influencia para la probabilidad de mala clasificación $P(i/j;R)$ puede derivarse de dos formas, uno usando la aproximación de la función de influencia, al reemplazar F por $\hat{F}_n$ y $\varepsilon = -\frac{1}{n-1}$, o también mediante la función de influencia general mostrada (28). Por ambos caminos de obtienen resultados similares.

a. Mediante la aproximación de la función de influencia

Tomando las consideraciones realizadas para expresar (27), donde se muestra la perturbación de la función de influencia muestral; podemos obtener la perturbación $P(i/j;R)$, entonces se tiene los siguiente: $\hat{\tilde{F}}^{(i)}_{n-1} = \phi\left(-\frac{1}{2}D_{(i)}\right)$, evaluando en la función de influencia de (26) se tiene:

$$\hat{I}\left(\vec{x}, MP\right) = \frac{\phi\left(-\frac{1}{2}D_{(i)}\right) - \phi\left(-\frac{1}{2}D\right)}{-\frac{1}{n_1 - 1}}$$

$$\hat{I}\left(\vec{x}, MP\right) = -(n_1 - 1)\left(\phi\left(-\frac{1}{2}D_{(i)}\right) - \phi\left(-\frac{1}{2}D\right)\right)$$

Finalmente, la medida de influencia para la probabilidad de mala clasificación es expresada como:

$$\hat{I}\left(\vec{x}, MP\right) = (n_1 - 1)\left[\phi\left(-\frac{1}{2}D\right) - \phi\left(-\frac{1}{2}D_{(i)}\right)\right] \qquad (37)$$

A continuación se presenta otra forma de derivar la medida de influencia para la probabilidad de mala clasificación.

b. Mediante la función de influencia general

Sea $\vec{X}_1 \vec{X}_2, \vec{X}_3, ..., \vec{X}_n$ una muestra aleatoria procedente de una vector aleatoria $\vec{X}$ p-dimensional, llevando en cuenta que la estimación de la probabilidad de mala clasificación es definida como $P(i/j;R) = \phi\left(-\frac{1}{2}D\right)$, para la construcción de la medida se toman las siguientes consideraciones:

- La estadística a tomar en cuenta para la deducción de la medida es la probabilidad de mala clasificación.
- La probabilidad de mala clasificación calculada sin la i-ésima observación es de denotada como $P(i/j;R)^{(i)} = \phi\left(-\frac{1}{2}D_{(i)}\right)$

Teniendo en cuenta la función de influencia general presentada en (28), la medida de influencia para la probabilidad de mala clasificación es dado como:

$$\hat{I}\left(\vec{x}_i; MP\right) = -(n_1 - 1).\left[\phi\left(\frac{D_{(i)}}{2}\right) - \phi\left(\frac{D}{2}\right)\right] \qquad (38)$$

donde $\phi(.)$ es la función de distribución normal estándar acumulada y $D_{(i)}$ es el estimador máximo verosímil de Δ eliminando la i-ésima observación de la muestra.

Según los resultados obtenidos en (37) y (38) podemos decir efectivamente por ambos caminos se obtienen resultados similares.

3.4.3. Medida de influencia alternativa para la probabilidad de mala clasificación

La medida que se presentará a continuación es una variante de la medida de influencia para la probabilidad de mala clasificación mostrada en (37) y (38) basado en lo siguiente:
La estimación de la probabilidad de mala clasificación para (37) y (38) puede ser obtenida bajo dos reglas de clasificación, la primera usando la usual regla de clasificación de Fisher (1936) mostrada en (1.2.3); la segunda mediante una regla construida a partir de la omisión de la i-ésima observación, que por motivos de cálculos asumiremos que la observación omitida es del grupo G_1.

Sea $\vec{X}_1, \vec{X}_2, \vec{X}_3, ..., \vec{X}_{n_k}$; $k = 1,2$ y $n_1 + n_2 = n$, muestras aleatorias de la variable aleatoria $\vec{X}$ p-dimensional, asumiendo que los estimadores de los vectores de medias y de la matriz de varianzas y covarianzas son $\overrightarrow{\overline{X}}_{(i)}^{(k)}$ as S_u respectivamente, se plantea el siguiente modelo en términos muestrales:

$$Y = \hat{\vec{\alpha}}_{(i)}^{T} \vec{x} \qquad (39)$$

donde: $\hat{\vec{\alpha}}_{(i)} = S_u^{-1}.\left(\overrightarrow{\overline{X}}_{(i)}^{(1)} - \overrightarrow{\overline{X}}^{(2)} \right)$

Teniendo en cuenta las consideraciones tomadas, para derivar la regla de clasificación mostrada en (1.2.3); se tiene la siguiente regla de clasificación, para el modelo planteado en (39):

Asignar $\vec{x}$ al grupo G_1 si:

$$\hat{\vec{\alpha}}_{(i)}^{T}\left[\vec{x} - \frac{1}{2}\left(\overrightarrow{\overline{X}}_{(i)}^{1} - \overrightarrow{\overline{X}}^{(2)} \right) \right] > 0 \qquad (40)$$

Caso contrario asignar al grupo G_2

Para derivar las probabilidades de mala clasificación bajo esta regla, se tienen las siguientes consideraciones:

En términos muestrales el vector aleatorio $\vec{X}$ p-dimensional tiene distribución $N_p\left(\overrightarrow{\overline{X}}_{(i)}^{(1)}; S_u\right)$.

De la combinación lineal definida en (39), se puede deducir entonces lo siguiente:

$$Y = \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} x \approx N_p\left(\overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} \overrightarrow{\overline{X}}_{(i)}^{(1)}; G^2\right)$$

donde, $G^2 = \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} S_u \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}$

En efecto, para el vector de medias

$$E(Y) = E\left[\overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} \vec{x}\right]$$

$$= \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} E\left[\vec{X}\right]$$

$$= \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} . \overrightarrow{\overline{X}}_{(i)}^{(1)}$$

Para la matriz de varianzas y covarianzas

$$V(Y) = V\left[\overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} \vec{x}\right]$$

$$= \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} . V\left[\vec{x}\right] . \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}$$

$$= \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} . S_u . \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T}$$

De este modo se tiene que $Y = \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} x \approx N\left(\overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} . \overrightarrow{\mu}^{(k)}; G^2\right)$

donde, $G^2 = \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T} S_u \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}$

Teniendo en cuenta que la observación omitida proviene del grupo G_1, se tiene las probabilidades de mala clasificación en el nuevo espacio.

Cuando $\vec{x}$ proviene del grupo G_1

$$P_{(i)}^{(1)} = P\left[\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\left[\vec{x} - \frac{1}{2}\left(\vec{\overline{X}}_{(i)}^{(1)} + \vec{\overline{X}}^{(2)}\right)\right] < 0 \,/\, \vec{X} \in G_1\right]$$

$$P_{(i)}^{(1)} = P\left[\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{x} - \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}_{(i)}^{(1)} - \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(2)} < 0 \,/\, \vec{X} \in G_1\right]$$

$$P_{(i)}^{(1)} = P\left[\frac{\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{x} - \vec{\mu}_Y}{\sigma_Y} < \frac{\frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}_{(i)}^{(1)} + \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(2)} - \overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}_{(i)}^{(1)}}{G}\right]$$

Aplicando propiedades elementales y aumentando y quitando el término $\frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(1)}$ al numerador del segundo miembro, se tiene:

$$P_{(i)}^{(1)} = P\left[z < \frac{\frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(2)} - \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}_{(i)}^{(1)} + \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(1)} - \frac{1}{2}\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\vec{\overline{X}}^{(1)}}{G}\right]$$

$$P_{(i)}^{(1)} = P\left[z < \frac{-\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\left(\vec{\overline{X}}^{(2)} - \vec{\overline{X}}^{(1)}\right) - \overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}_{(i)}^{(1)}\right)}{2G}\right]$$

$$P_{(i)}^{(1)} = \phi\left[\frac{-\overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}^{(2)}\right) - \overset{\wedge}{\vec{\alpha}}_{(i)}^{T}\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}_{(i)}^{(1)}\right)}{2G}\right]$$

De manera similar, se puede obtener $P_{(i)}^{(2)}$

$$P_{(i)}^{(2)} = \phi\left[\frac{-\overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T}\left(\overset{\rightarrow}{\overline{X}}^{(1)} - \overset{\rightarrow}{\overline{X}}^{(2)}\right) + \overset{\wedge}{\underset{\rightarrow}{\alpha}}_{(i)}^{T}\left(\overset{\rightarrow}{\overline{X}}^{(1)} - \overset{\rightarrow}{\overline{X}}_{(i)}^{(1)}\right)}{2.G}\right]$$

donde $\alpha_{(i)} = S_1\left(\overset{\rightarrow}{\overline{X}}_{(i)}^{(1)} - \overset{\rightarrow}{\overline{X}}^{(2)}\right);\ G^2 = \overset{\wedge}{\alpha}_{(i)}^{T} S \overset{\wedge}{\alpha}_{(i)}$

Fung (1995), propone una medida similar a la mostrada en (38), pero diferente forma de estimar las probabilidades de mala clasificación, como:

$$DMP_i = \left[\frac{1}{2}\left(P_{(i)}^{(1)} + P_{(i)}^{(2)}\right)\right] - \left[\phi\left(-\frac{1}{2}D\right)\right] \qquad (41)$$

Una forma sencilla de mostrar (41), es aproximando las probabilidades de mala clasificación $P_{1(i)}$ y $P_{2(i)}$ al segundo orden de la serie de Taylor alrededor de $-\frac{1}{2}D$, se demuestra que la medida DMP_i, puede ser aproximadamente igual a

$$DMP_i \cong \frac{\phi\left(-\frac{1}{2}.D\right)}{4.D.(n_1-1)^2}.\left[\left(1 - w_k.\overset{\wedge}{\psi}_i\right)^2.\left(d_i^2 - \frac{\overset{\wedge}{\psi}_i^{\,2}}{D^2}\right) + \frac{1}{4}\overset{\wedge}{\psi}_i^{\,2}\right] \qquad (42)$$

Donde: $d_i^2 = \left(\vec{x}_i^{(k)} - \overset{\rightarrow}{\overline{X}}^{(k)}\right)^T S_u.\left(\vec{x}_i^{(k)} - \overset{\rightarrow}{\overline{X}}^{(k)}\right)$

$$\overset{\wedge}{\psi}_i = \overset{\rightarrow}{\overset{\wedge}{\alpha}}^{T}.\left(\vec{x}_i^{(k)} - \overset{\rightarrow}{\overline{X}}^{(k)}\right)^T$$

3.4.4. Medida de influencia para las puntuaciones de la función discriminante

Fung (1995), propuso una medida para las puntuaciones de la función discriminante de Fisher, siguiendo básicamente la metodología propuesta por Cook y Weisberg (1982)[7], basada en cuantificar el efecto que tiene la omisión de una observación en el vector de parámetros, para ello, se basó en la relación de equivalencia entre los coeficientes de la función discriminante lineal de Fisher y el modelo del análisis de regresión lineal múltiple, para la deducción de dicha medida se toma en cuenta las siguientes consideraciones.

- $\hat{\vec{\alpha}}^T \vec{x} - \frac{1}{2}\hat{\vec{\alpha}}^T \left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}^{(2)} \right)$, son las puntuaciones de la función discriminante, que también pueden ser denotados como

  $\vec{\beta}^T \vec{x}$, donde

  $\vec{\beta}^T = \left[-\frac{1}{2}\hat{\vec{\alpha}}^T \left(\vec{\overline{X}}^{(1)} + \vec{\overline{X}}^{(2)} \right),\ \hat{\vec{\alpha}}^T \right]$ y $\vec{x}^T = \left[1\ ,\ \vec{x}^T \right]$

- Interesa el efecto de la omisión de la observación i (por simplicidad, asumimos que proviene de π_1) sobre el parámetro $\vec{\beta}$, al que denotamos $\vec{\beta}_{(i)}$.
- Dicho efecto se evalúa a través de la diferencia de las puntuaciones de la función discriminante, con y sin la i-ésima observación $\vec{\beta}^T \vec{x} - \vec{\beta}_{(i)}^T \vec{x}$
- Johnson[8] (1987), bajo el marco bayesiano propuso las siguientes medidas,

$E\left[\vec{\beta}^T \vec{x} - \vec{\beta}_{(i)}^T \vec{x} \right]$ y $E\left| \vec{\beta}^T \vec{x} - \vec{\beta}_{(i)}^T \vec{x} \right|$

[7] Referencia en Fung (1995), Enguix (2001)
[8] Referencia en Fung (11)

La primera medida no sería tan adecuada, porque pasaría cerca a cero, debido a las cancelaciones de los efectos individuales durante la integración o la adición la otra porque no puede ser resumido en términos de las estadísticas fundamentales.

- Fung. Propuso la siguiente medida:

$$E\left[\overset{\rightarrow}{\beta}^{T}\vec{x}-\overset{\rightarrow}{\beta}_{(i)}^{T}\vec{x}\right]^{2},$$

el promedio de las diferencias al cuadrado de las puntuaciones de la función discriminante con la muestra completa y con la muestra sin la observación i, y asumiendo que $\vec{X}$ es distribuido como:

$$t.N\left(\mu^{(1)},\Sigma\right)+\left(1-t\right)N\left(\mu^{(2)},\Sigma\right).$$

Tomando las estimaciones de los parámetros $\overline{\vec{X}}^{(1)}$, $\overline{\vec{X}}^{(2)}$ S_u y $t=\frac{n_1}{n}$, la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher, se define como:

$$E2=t.\beta_1^2+\left(1-t\right).\beta_2^2+V \qquad (43)$$

donde: $\beta_1=\dfrac{\left(\overset{\rightarrow}{\hat{\alpha}}-\overset{\rightarrow}{\hat{\alpha}}_{(i)}\right)^{T}\left(\overline{\vec{X}}^{(1)}-\overline{\vec{X}}^{(2)}\right)}{2}-\dfrac{\overset{\rightarrow}{\hat{\alpha}}_{(i)}\left(\overline{\vec{X}}^{(1)}-\vec{x}_{(i)}^{(1)}\right)}{2}$

$$\beta_2=\frac{-\left(\overset{\rightarrow}{\hat{\alpha}}-\overset{\rightarrow}{\hat{\alpha}}_{(i)}\right)^{T}\left(\overline{\vec{X}}^{(1)}-\overline{\vec{X}}^{(2)}\right)}{2}-\frac{\overset{\rightarrow}{\hat{\alpha}}_{(i)}\left(\overline{\vec{X}}^{(1)}-\vec{x}_{(i)}^{(1)}\right)}{2}$$

$$V=\left(\overset{\rightarrow}{\hat{\alpha}}-\overset{\rightarrow}{\hat{\alpha}}_{(i)}\right)^{T}.S_u.\left(\overset{\rightarrow}{\hat{\alpha}}-\overset{\rightarrow}{\hat{\alpha}}_{(i)}\right)$$

3.4.5. Medidas de influencia adicionales en el análisis discriminante lineal

Fung (1992), no solo propuso la función de influencia alternativa para las probabilidades de mala clasificación y para las puntuaciones de la función discriminante, si no también sugirió utilizar las estadísticas d_i^2 y $\hat{\psi}_i^{(k)}$ que están presentes en la construcción de la mayoría de medidas de influencia en el análisis discriminante lineal.

a. Primera medida

Esta medida, es la diferencia entre las puntuaciones de la función discriminante lineal, y el promedio discriminante de cada grupo y es expresado de la siguiente forma:

$$\hat{\psi}_i^{(K)} = \hat{\vec{\alpha}}^T \left(\vec{x}_i^{(k)} - \vec{\overline{X}}^{(k)} \right) = \hat{\vec{\alpha}}^T \vec{x}_i^{(k)} - \hat{\vec{\alpha}}^T \vec{\overline{X}}^{(k)} \qquad (44)$$

donde: $i = 1;2;3,...n_k ; k = 1,2.$

La medida mostrada, servirá para detectar las observaciones discordantes en cada grupo.

Expresa la distancia de cada observación respecto al centro de la población que se está analizando, y estas distancias están ponderadas por los coeficientes de la función discriminante lineal de Fisher.

b. Segunda medida

Esta medida es usada con frecuencia para detectar observaciones influyentes en el análisis de regresión. En el caso del análisis discriminante se usará para detectar observaciones influyentes en cada grupo y es definida como:

$$d_i^{(k)2} = \left(\vec{x}_i^{(k)} - \vec{\overline{X}}^{(k)} \right)^T S_u^{-1} \left(\vec{x}_i^{(k)} - \vec{\overline{X}}^{(k)} \right) \qquad (45)$$

donde: $i = 1;2;3,...n_k ; k = 1,2.$

La medida mostrada en (41), mide la distancia entre cada observación y su centro y estas distancias están ponderadas con la inversa de la matriz de covarianzas conjunta.

3.5. VALIDACIÓN DE LAS MEDIDAS DE INFLUENCIA

Se presenta a continuación un procedimiento que permitirá, evaluar el potencial de las medidas de influencia propuestas en esta sección, basada en la sugerencia dada por Beckman y Cook (1983).

- Se simulará mediante el método de Monte Carlo, dos muestras con distribuciones normales multivariantes, con la condición de que las matrices sean Homocedásticas.

- Una vez obtenidas las muestras multivariantes, se perturba una observación seleccionada al azar. Esta perturbación según la literatura revisada puede hacerse de dos formas: Cambiando la observación seleccionada, por otra observación que tenga características diferentes a las demás observaciones pero a la vez que sea relevante o multiplicando la observación seleccionada, por alguna constante.

- Finalmente se espera que al evaluar las medidas de influencia, la observación perturbada, sobresalga frente a las demás.

La simulación de variables con densidad conjunta normal multivariada se fundamenta en la técnica de descomposición de Cholesky (Anderson, 1984), cuya metodología se describe a continuación. La simulación de los datos se ha realizado usando el software libre R. El programa se presenta en el apéndice C.

- Se genera el vector de medias $\hat{\mu}_z$

- Se genera una o la matriz de varianzas y covarianzas $\hat{\Sigma}_z$.

- A esta matriz de covarianzas , se le aplica la descomposición de Cholesky:
  $\Sigma_z = L * L^T$, donde L^T es una matriz triangular superior

- Usando el teorema referente a la distribución de combinaciones lineales de vectores con distribución normal multivariada (Anderson, 1984), el vector simulado se expresa como:

$$\vec{X} = \hat{\vec{\mu}}_z + L * \vec{\varepsilon}$$, donde $\vec{\varepsilon} \approx N(0; I)$.

CAPITULO IV

APLICACIÓN DE LAS MEDIDAS DE INFLUENCIA

4.1. INTRODUCCIÓN

La ilustración de las medidas de influencia se realizó a través de un conjunto de datos simulados con distribución normal multivariante y tres conjuntos de datos reales, los mismos que en el primer capítulo se usaron para comprobar la relación existente entre los coeficientes de la función lineal discriminante y de la función de regresión lineal múltiple.

Las medidas de influencia fueron desarrolladas usando el software MATLAB, cuyo programa se adjunta en el apéndice D.

4.2. APLICACIONES

4.2.1. Caso: Datos simulados

- Se generaron los vectores de medias de cada uno de los grupos, habiéndose elegido los siguientes vectores:

$$\vec{\bar{X}}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \qquad \text{y} \qquad \vec{\bar{X}}^{(2)} = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

- Con sentencias del software R se generaron las matrices de varianzas y covarianza, a las que se les aplicó la descomposición de Cholesky.
- Usando el teorema referente a la distribución de combinaciones lineales se generó la matriz de datos que se presenta en el apéndice B.
- Se verificaron los supuestos de normalidad de los datos simulados, a través de la prueba de asimetría y de curtosis(Mardia, 1979).

- Mediante la prueba M de Box (Mardia, 1979), se verificó que las muestras proceden de poblaciones con iguales estructuras de varianzas y covarianzas.

- Aleatoriamente de seleccionó la observación 10 que tenía como datos originales a

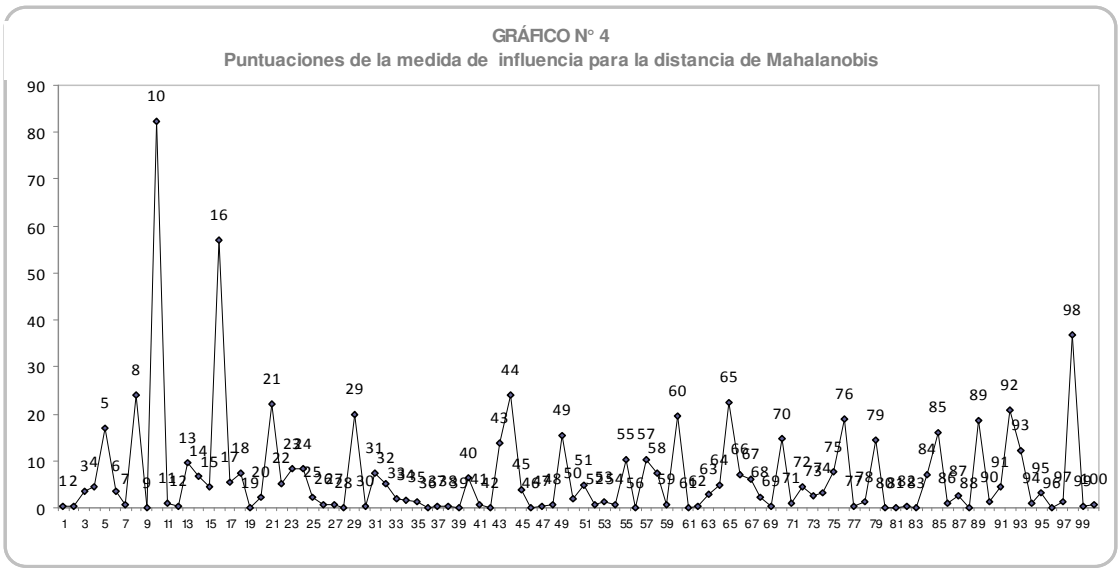
$$\overrightarrow{x_{10}}^{(1)} = \begin{bmatrix} 0.52 \\ 4.08 \\ 0.71 \\ -0.86 \end{bmatrix}$$

Utilizando en método de perturbación mencionada anteriormente, se cambió esta observación por la siguiente:

$$\overrightarrow{x_{10}}^{(1)} = \begin{bmatrix} 2.60 \\ 20.4 \\ 3.55 \\ -4.30 \end{bmatrix}$$

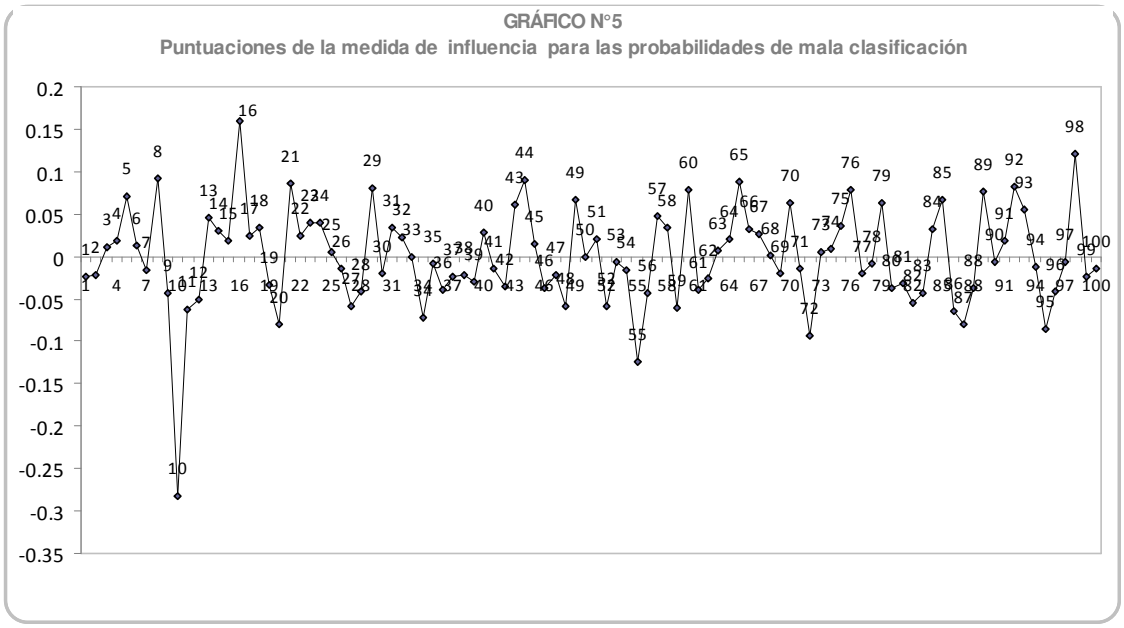
- Las puntuaciones usando las medidas según las fórmulas (36), (37), (38), (41) y (42) se presentan en el Apéndice C.
- Se muestra las correspondientes representaciones gráficas que involucran a las observaciones versus valor de la correspondiente medida según fórmulas (36), (37), (38), (41) y (42)

Según la fórmula (36) que es la medida de influencia para la distancia de Mahalanobis, representado en el GRÁFICO Nº 4, la observación 10 es la influyente, que es precisamente la observación que se había perturbado.

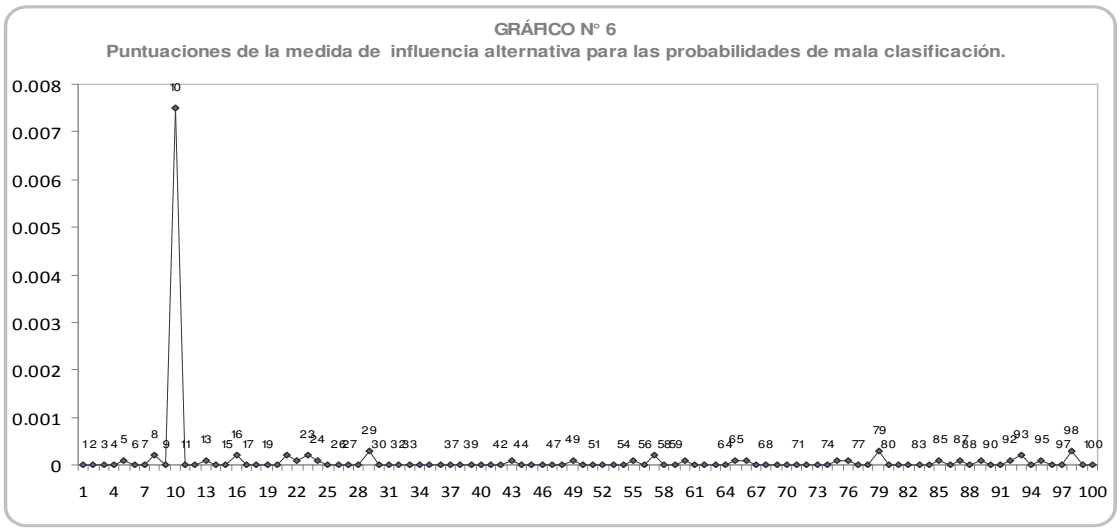


En el GRÁFICO Nº 5, está la representado las puntuaciones de la medida de influencia para la probabilidad de mala clasificación, con fórmula (37), la observación perturbada (la observación Nº 10) es la sobresale frente a todas las demás, aun que esta vez es menor a todas, y esto es justificable debido a la forma que tiene esta medida,

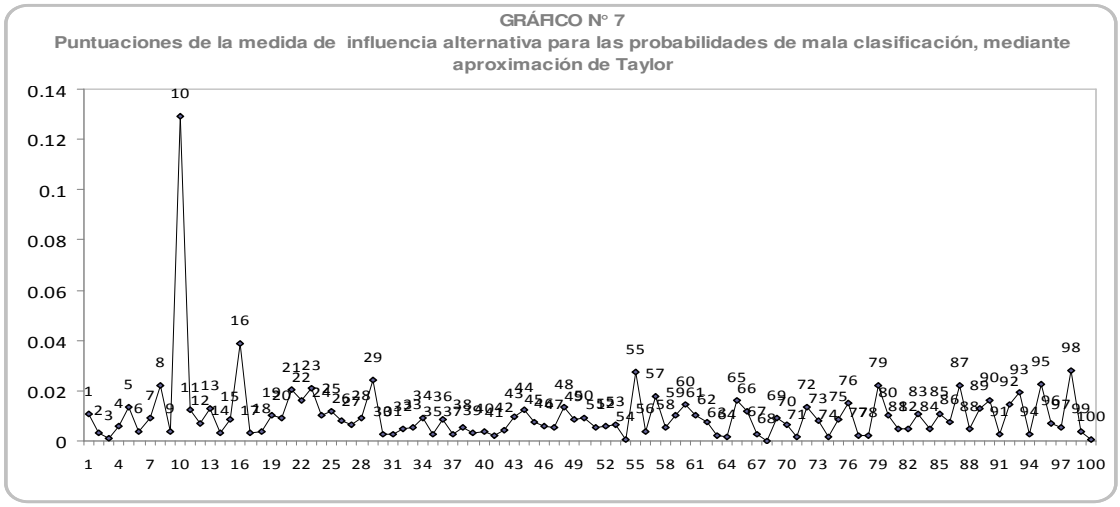
$\hat{I}\left(\vec{x}_i; MP\right) = -(n_1 - 1).\left[\phi\left(\frac{D_{(i)}}{2}\right) - \phi\left(\frac{D}{2}\right)\right]$, pues si $\phi\left(\frac{D_{(i)}}{2}\right)$ es mayor a $\phi\left(\frac{D}{2}\right)$, $\hat{I}\left(\vec{x}_i; MP\right)$ será negativa.



En el GRÁFICO N° 6, se muestra la representación gráfica de la medida de influencia alternativa para la probabilidad de mala clasificación, con fórmula (38), medida alternativa a la medida anterior (medida de influencia para la probabilidad de mala clasificación), debido a que esta se construye a partir de una nueva regla de clasificación, cada vez que una observación sea omitida, deducida en 3.4 ítem c). Esta medida, también identifica a la observación perturbada (la observación número 10) pero de una manera contundente.



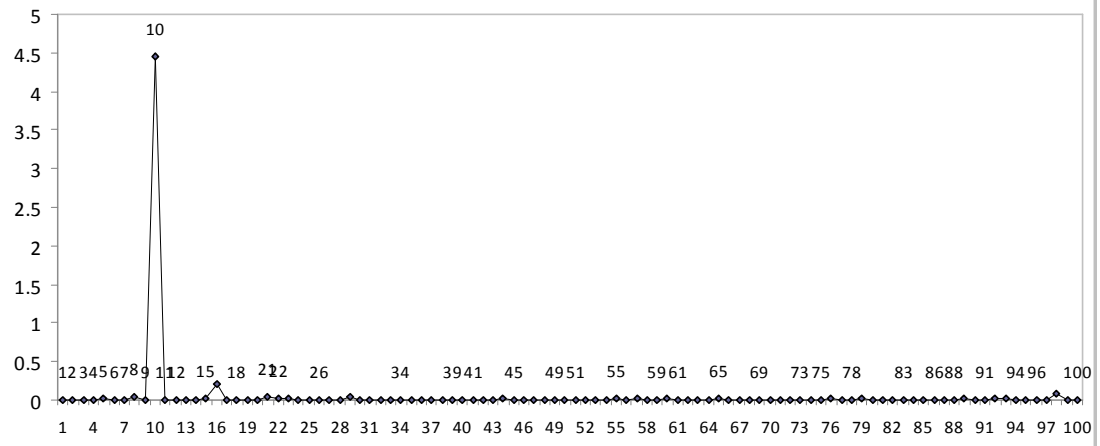
Las puntuaciones de la medida de influencia para la probabilidad de mala clasificación usando aproximación de Taylor, según fórmula (42), se muestra en el GRÁFICO N° 7 , igualmente, la observación N° 10 es la que sobresale.



En el GRÁFICO N°8, se presentan las puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher según fórmula (43).

GRÁFICO Nº 8

Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher.



Las representaciones gráficas de las puntuaciones de las diferentes medidas presentadas, confirman que la observación Nº 10 que fue perturbada, es la que sobresale frente a las demás, por lo tanto será considerada como influyente. Hay que resaltar que la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de Taylor y la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher son las que ofrecen resultados muy similares y contundentes, sustentada con una correlación de 0.997 entre las puntuaciones de ambas medidas; la medida para la probabilidad de mala clasificación y la medida alternativa para la probabilidad de mala clasificación usando la aproximación de Taylor también ofrecen resultados muy similares debido a que la correlación entre ambas es de 0.899. Finalmente podemos tener la certeza que las distintas medidas de influencia propuestas en el capítulo anterior, identifican a las observaciones influyentes.

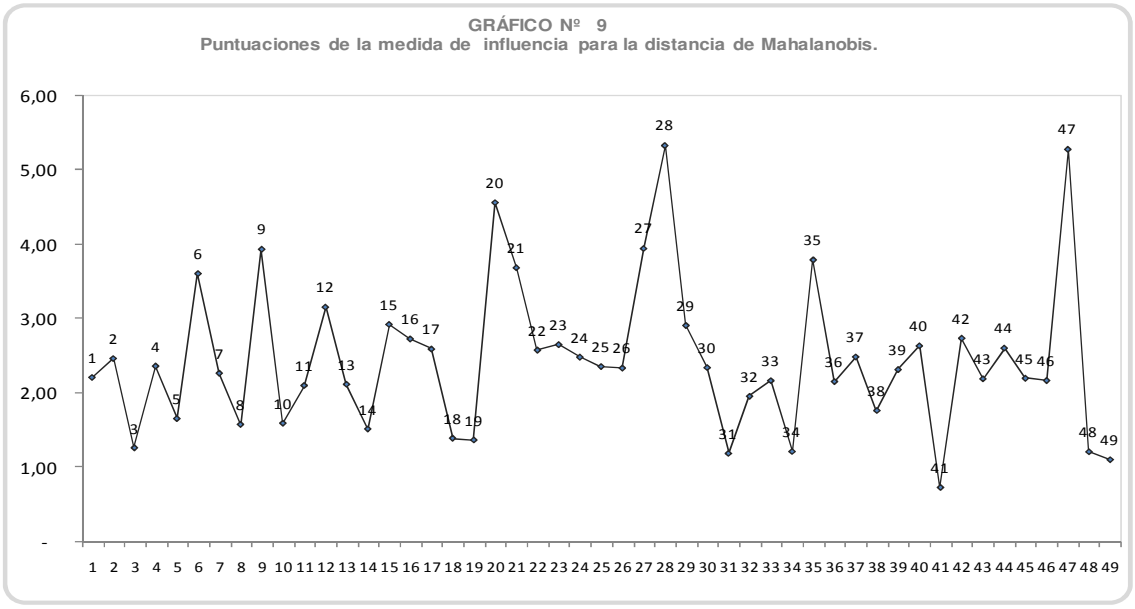
4.2.2. Caso: Primer conjunto de datos

Como ya se mencionó en el primer capítulo, este conjunto de datos corresponde a un grupo de gorriones moribundos que fueron estudiados en un laboratorio de Brown University, Rhode Island en febrero de 1898(Manly, 2005). Este conjunto de datos, cumple con el supuesto de homocadásticidad de las matrices de varianzas y covarianzas.

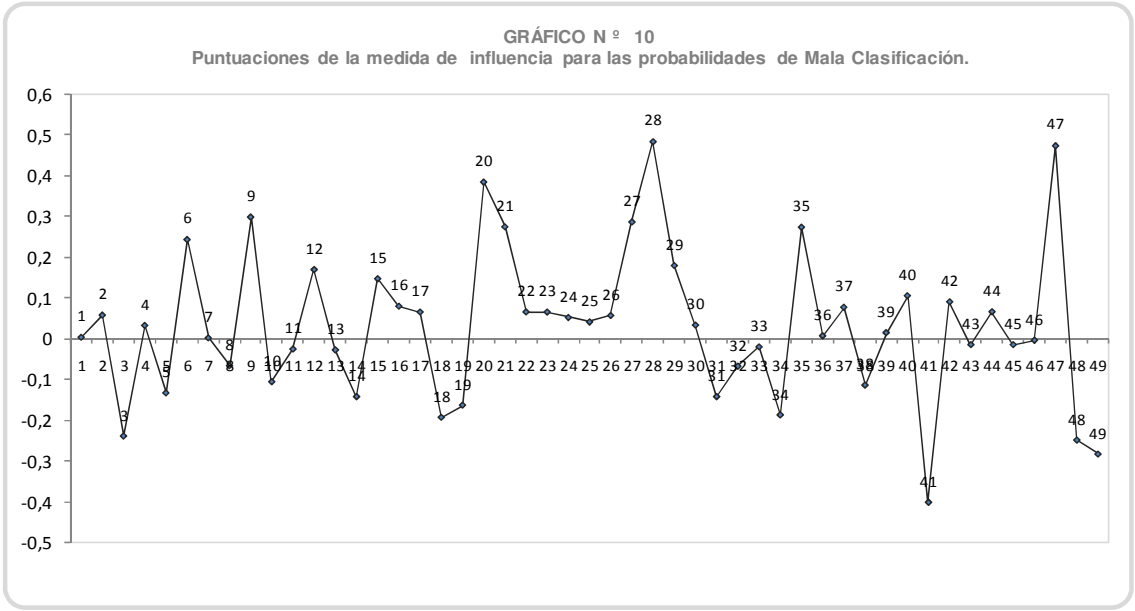
Las puntuaciones usando las medidas según las fórmulas (36), (37), (38), (41) y (42) se encuentran en el Apéndice C.

En el GRÁFICO Nº 4, se presentan las puntuaciones de la medida de influencia para la distancia de Mahalanobis, según fórmula (36), donde las observaciones que son potencialmente

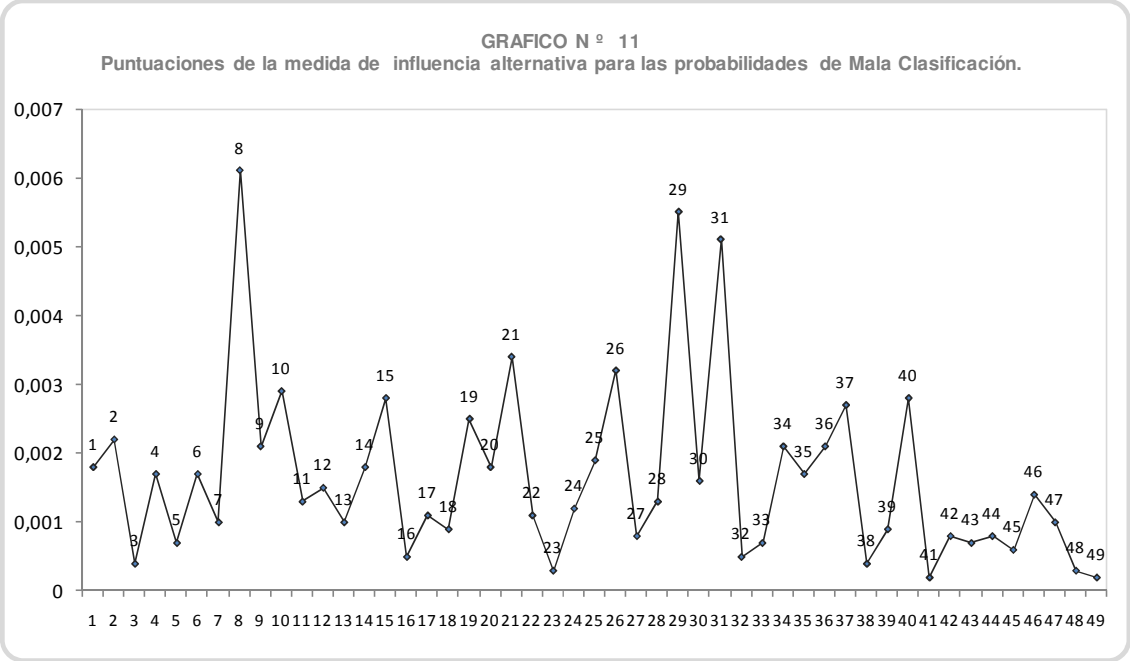
influyentes son la 47, 28, 20, 27 y 9. Cabe indicar que todas estas observaciones fueron mal clasificadas según la función discriminante lineal de Fisher.



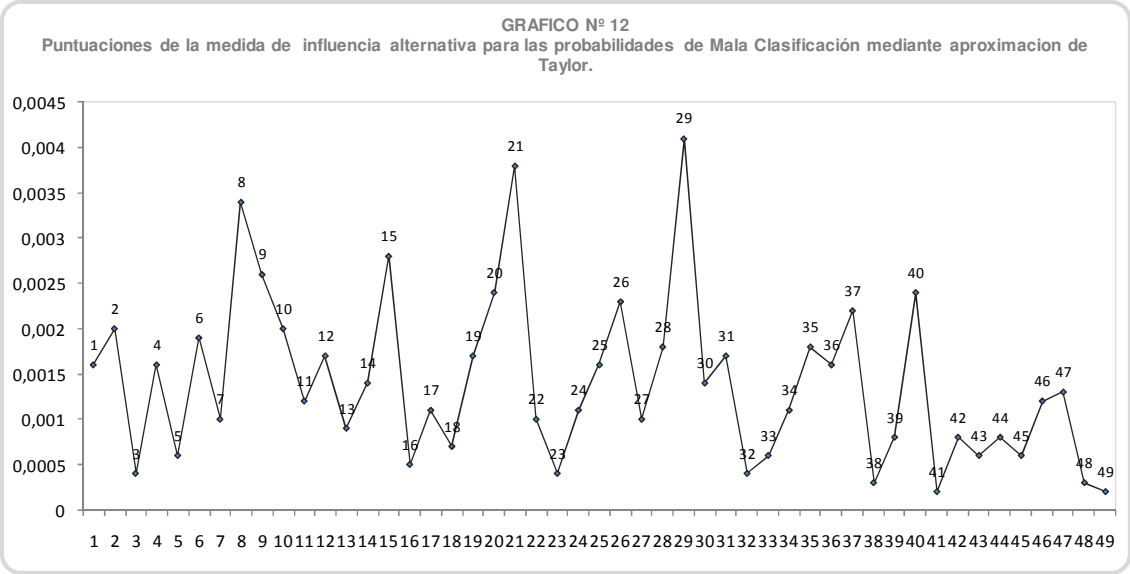
En el GRÁFICO Nº 10, se muestra las puntuaciones de la medida de influencia para la probabilidad de mala clasificación, según fórmula (37), en ella puede observarse que las observaciones 28, 47, 20 del segundo grupo y 9 del primer grupo, son potencialmente influyentes, en general las puntuaciones de esta mediada son similares a las puntuaciones de la influencia para la distancia de Mahalanobis, esto es sustentado con una alta correlación entre ambas de 0.98.



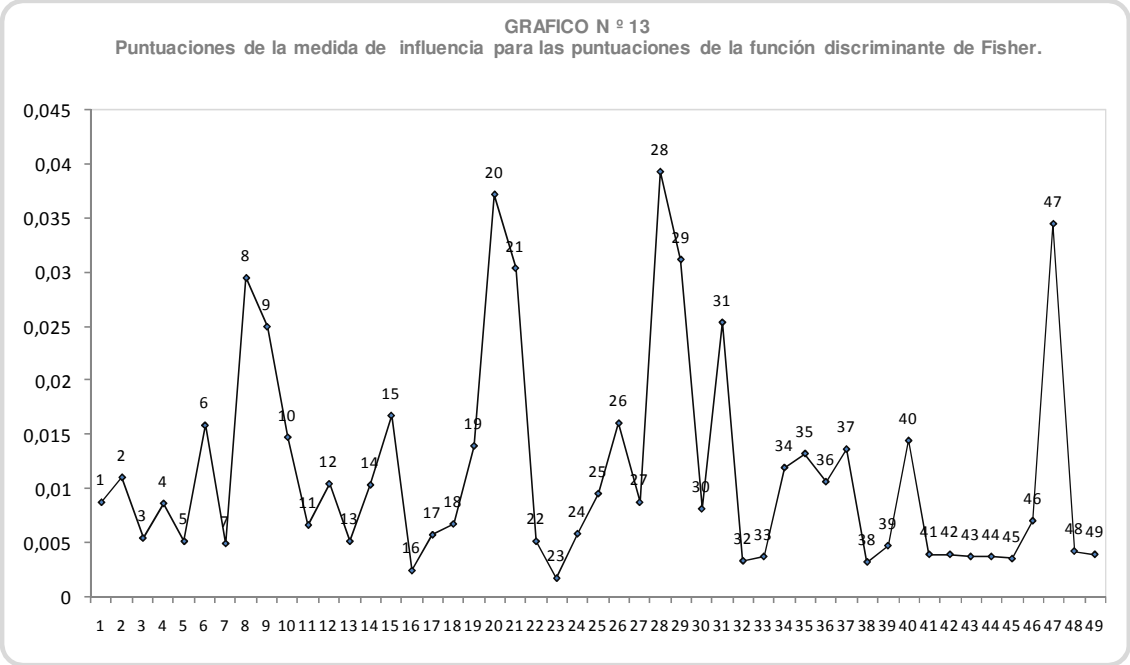
Las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación, según fórmula (38), se muestra en el GRÁFICO Nº 11, hay que recordar que esta medida es una variante de la medida para la probabilidad de mala clasificación, las observaciones que pueden considerarse influyentes son la 8, 29, 31 y la 21 (solo las observaciones 21 y 29 fueron mal clasificadas, según la la función discriminante lineal de Fisher).



La representación gráfica de las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de Taylor, según fórmula (42), se muestra en el GRÁFICO Nº 12, las observaciones que pueden considerarse como influyente son 29, 21 y la 8, resultados similares a las puntuaciones de la medida anterior (correlación entra ambas de 0.87).



En el GRÁFICO Nº 13, se muestra gráficamente la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher según, fórmula (43), las observaciones potencialmente influyentes son 28, 20, 47, 29, 21 y la 8.



En la TABLA N°4 se presentan las puntuaciones de las medidas de influencias adicionales para cada grupo, según la primera medida las observaciones potencialmente influyentes son la 20, 9 y la 21 esto en el primer grupo y la 28 y la 47 en el segundo grupo y según la segunda media adicional la 8 en el primer grupo y la 31 en el segundo grupo.

TABLA N° 4 Puntuaciones de las medidas de influencia adiconales 1/							
Primera Medida				Segunda Medida			
Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1
20	4.565	28	5.335	8	10.536	31	18.101
9	3.937	47	5.285	19	6.716	34	11.504
21	3.687	27	3.946	10	6.355	29	11.032
6	3.608	35	3.793	21	6.043	26	8.280
12	3.156	29	2.909	20	5.897	28	8.068
15	2.923	42	2.737	9	4.942	47	7.352
16	2.727	23	2.651	14	4.862	40	7.173
17	2.591	40	2.635	15	4.599	37	7.087
2	2.463	44	2.605	6	3.719	36	6.454
4	2.362	22	2.577	2	3.712	41	6.453

1/ solo se presentan las 10 mayores puntuaciones de cada grupo

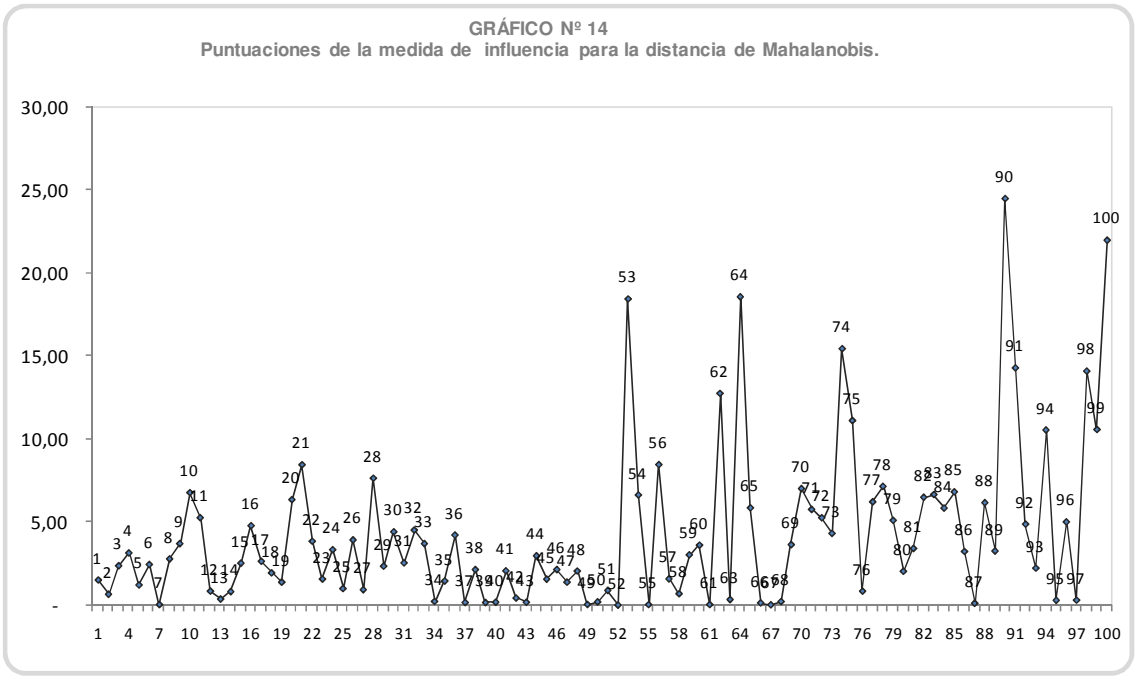
Teniendo en cuenta las puntuaciones de las diferentes medidas las observaciones que son consideradas como influyentes son la 28 y 47, pues la omisión de ellas, altera la estimación de los siguientes parámetros:

- La estimación de la distancia de Mahalanobis de 0.2353 a 0.3858 y a 0.3792 respectivamente.
- La estimación de las probabilidad de mala clasificación de 0.4042 a 0.3781 y a 0.3741 respectivamente.
- Mejora la Tasa de error aparente de 0.35 a 0.33 y a 0.2709.

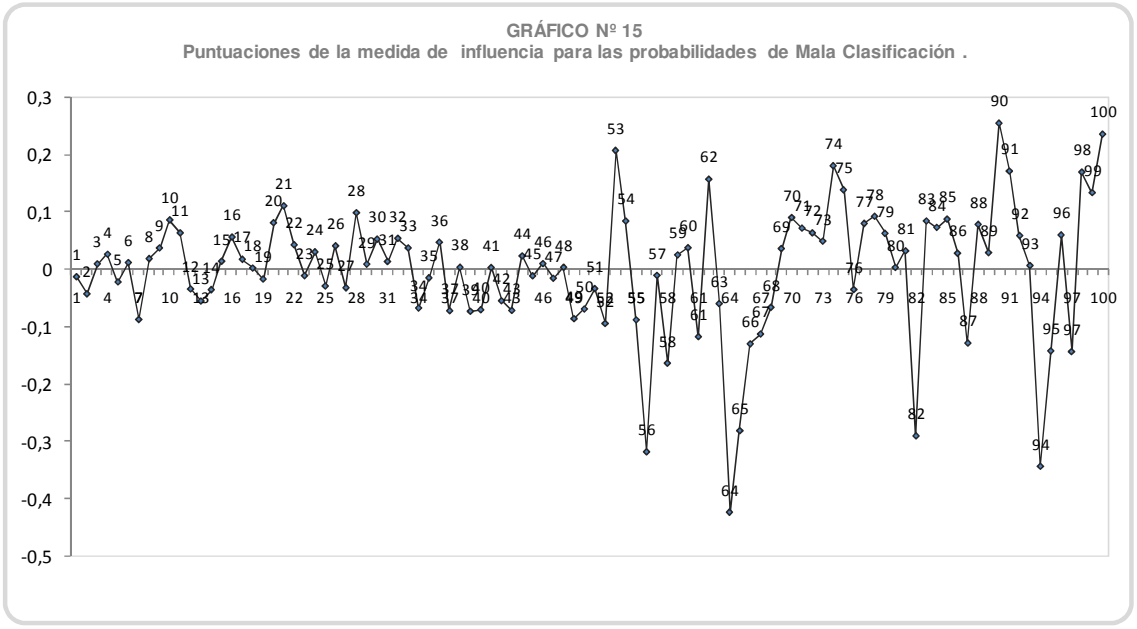
4.2.3. Caso: Segundo conjunto de datos

Este conjunto de datos(Gomez, et.al, 2008), también ya fue usado en el capítulo 1 para comprobar la relación de proporcionalidad entre los coeficientes de la función lineal discriminante y los coeficientes de la función de regresión lineal múltiple. Se comprobó también que las matrices de covarianzas no cumple con el supuesto de homocedasticidad.

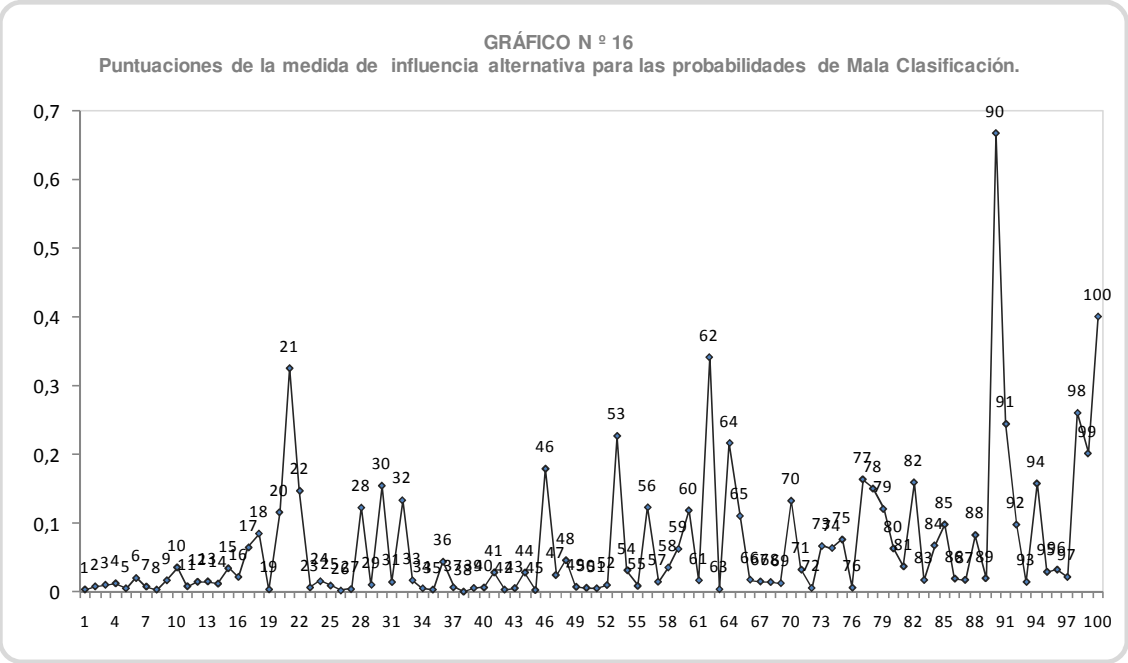
Las puntuaciones de la medida de influencia para la probabilidad de mala clasificación, según fórmula (36), se muestran en el GRÁFICO Nº 14, las observaciones que podrían considerarse como influyentes son la 90, 100, 64 y 53.



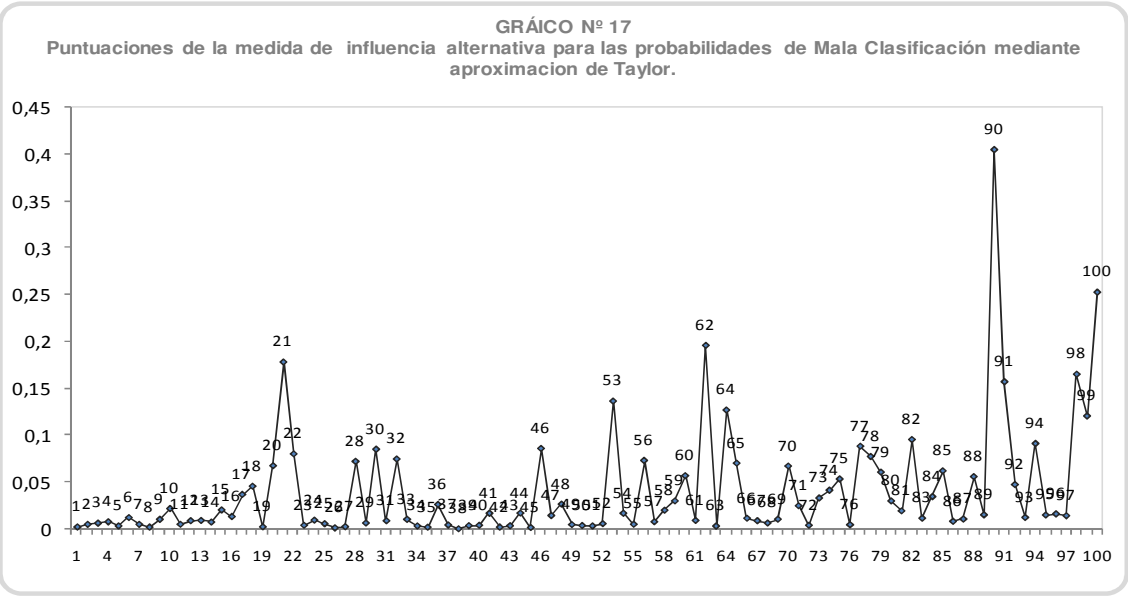
La representación gráfica de la medida de influencia para la probabilidad de mala clasificación, según fórmula (37), se muestra en el GRÁFICO Nº 15, según ella, las observaciones potencialmente influyentes son la 90, 100, 53 (puntuaciones positivas) y la 64,y 94 (puntuaciones negativas).



En el GRÁFICO Nº 16, se muestra las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación, según fórmula (38), las observaciones que potencialmente influyentes son la 100, 90, 62 y la 21.

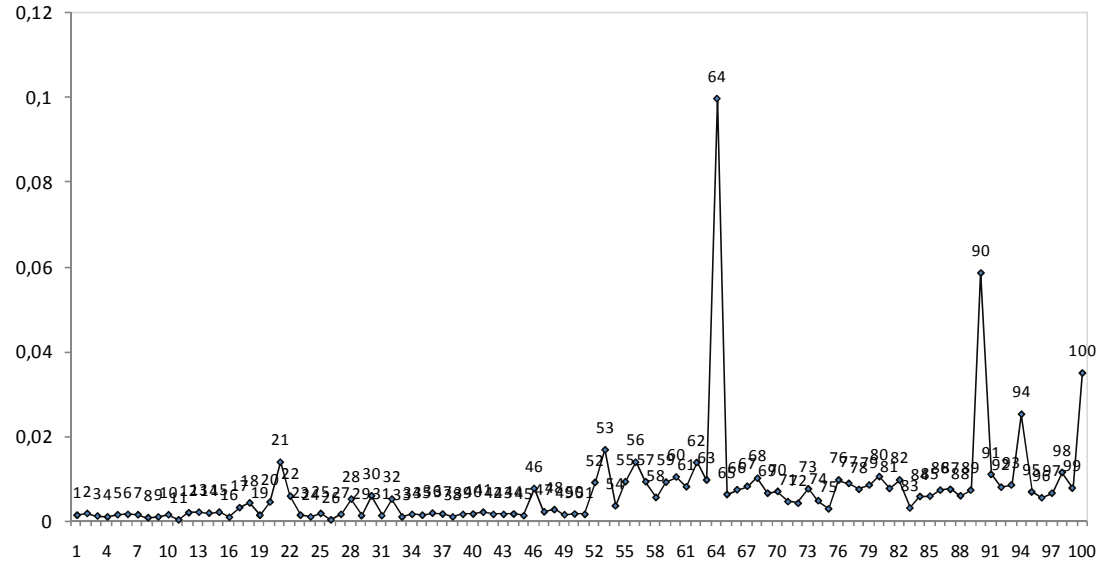


En el GRÁFICO Nº 17, se muestra las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de Taylor, según fórmula (42), las observaciones que potencialmente influyentes son la 100, 90, 62 y la 21, hay que mencionar que existe una correlación perfecta entre esta medida y la anterior (correlación 1.00).



Las puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher según, fórmula (43), se muestran en el GRÁFICO Nº 18, las observaciones potencialmente influyentes son 64, 90 y 100.

GRÁFICO Nº 18
Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante de Fisher.



Según las puntuaciones de las medidas de influencia adicionales para ambos grupos mostrada en la TABLA N°5, las observaciones potencialmente influyentes según la primera medida son la 21 y la 28 en el primer grupo la 90 y la 100 del segundo grupo y según la segunda medida la 46 del primer grupo y la 64 del segundo grupo.

TABLA N° 5
Puntuaciones de las medidas de influencia adiconales 1/

Primera Medida				Segunda Medida			
Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1
21	8.468	90	24.510	46	11.988	64	13.983
28	7.661	100	21.994	21	7.043	90	9.651
10	6.790	64	18.588	18	7.036	94	9.598
20	6.366	53	18.460	22	6.307	82	8.586
11	5.277	74	15.462	13	5.951	56	8.161
16	4.798	91	14.305	30	5.886	100	7.437
32	4.536	98	14.113	32	5.073	65	7.219
30	4.431	62	12.770	17	4.186	67	6.728
36	4.231	75	11.133	48	3.855	62	6.094
26	3.939	99	10.592	20	3.538	98	5.360

1/ solo se presentan las 10 mayores puntuaciones de cada grupo

Los resultados de las distintas medidas de influencia, indican que las observaciones influyentes son la 90, la 100 y la 64 pues la omisión de ellas, altera la estimación de los siguientes parámetros:

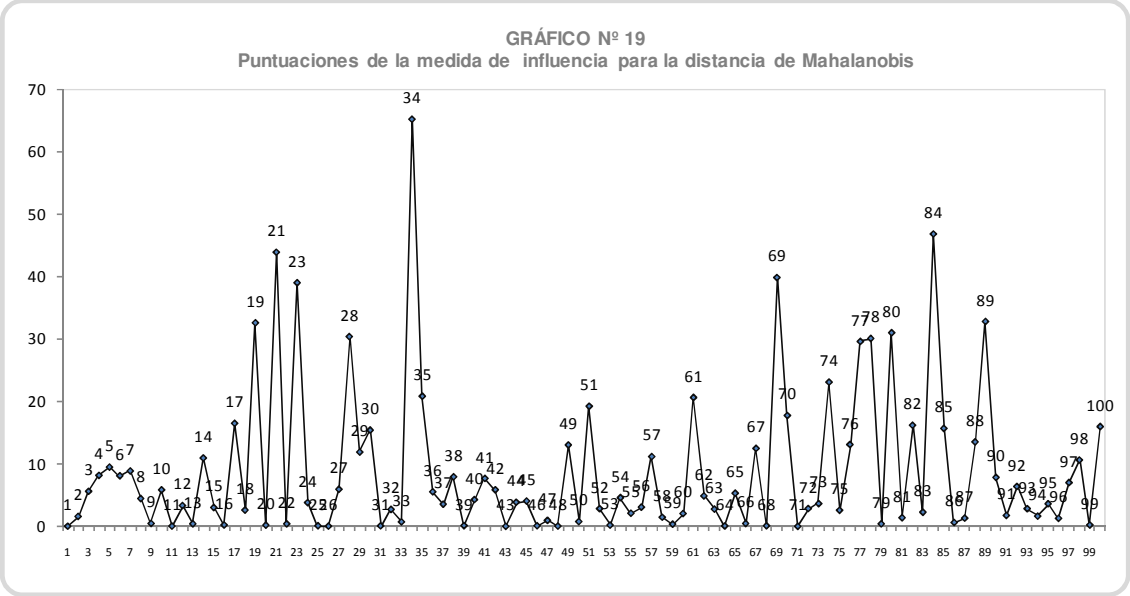
- La estimación de la distancia de Mahalanobis de 4.6508 a 5.1117 y 5.0498 y a 4.9960 respectivamente.
- Las estimaciones de las probabilidades de mala clasificación de 0.1405 a 0.1298; a 0.1306 y a 0.1319 respectivamente.

- Altera la tasa de error aparente de 8.0909 a 8.0808; y a 9.0909.

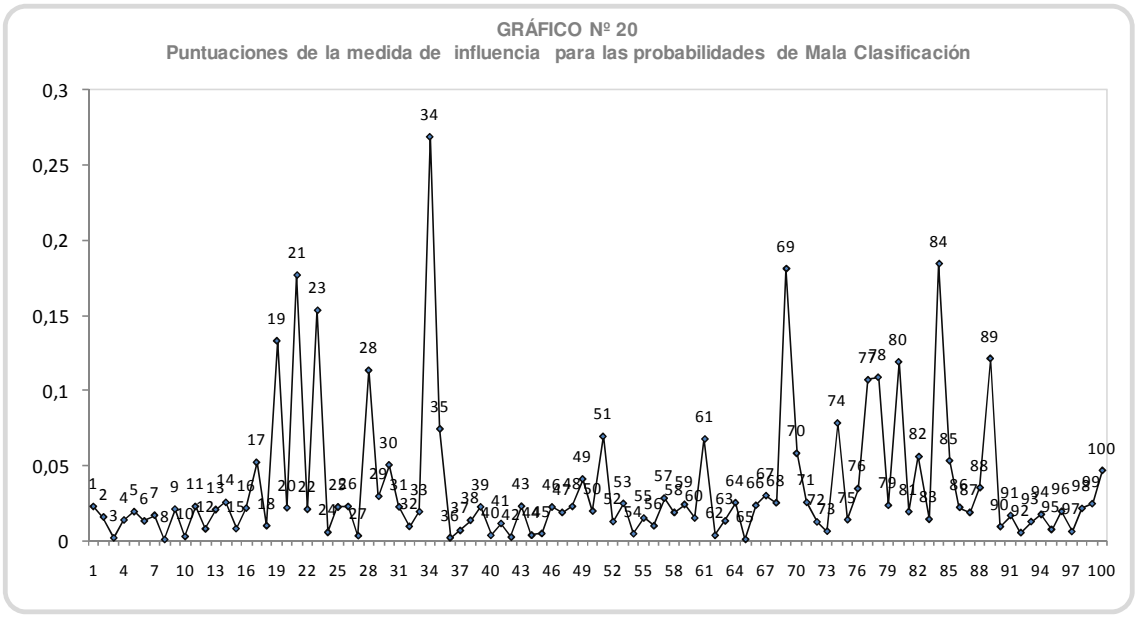
4.2.4. Caso: Tercer conjunto de datos

El tercer conjunto de datos corresponde a especies (versicolor y virginica) de iris, este conjunto de datos no cumple con el supuesto de homocedasticidad de las matrices de covarianzas, a continuación de muestra las representaciones gráficas de las distintas medidas y en el ANEXO C se encuentran las correspondientes puntuaciones.

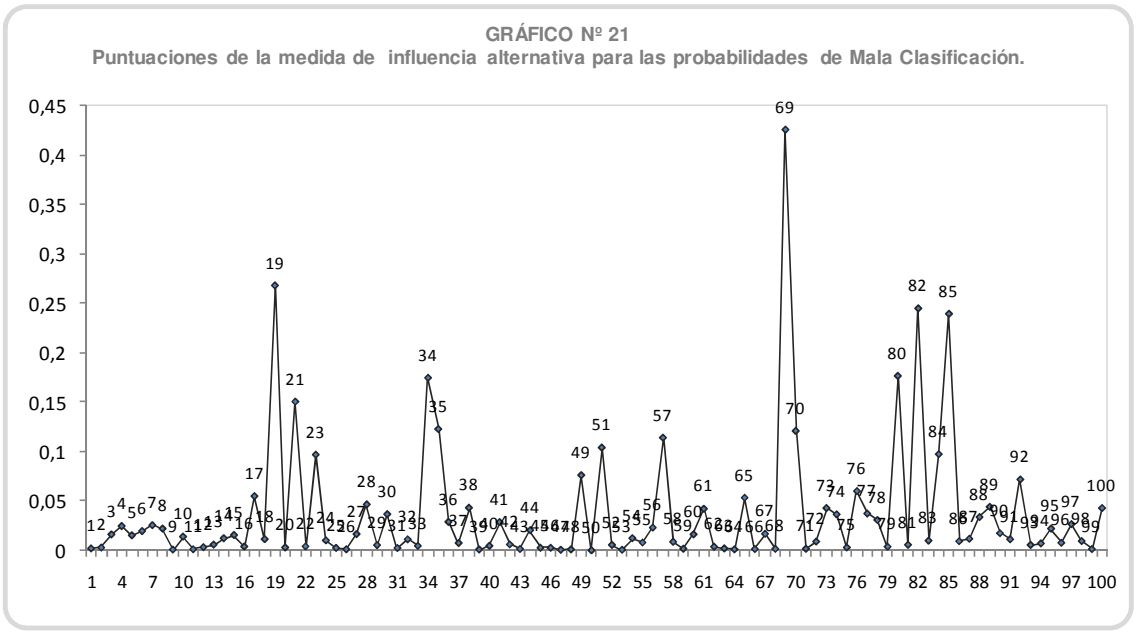
La representación gráfica de la medida de influencia para la distancia de Mahalanobis, según fórmula (36), se muestra en el GRÁFICO Nº 19, las observaciones que pueden considerarse como influentes son la 34, 84, 21 y 19 (de este grupo, solo la observación 19 fue bien clasificada la función discriminante lineal de Fisher).



En el GRÁFICO Nº 20, se muestra gráficamente las puntuaciones de la medida de influencia para la probabilidad de mala clasificación, según fórmula (37), las observaciones que pueden considerarse como influyentes son la 34, 84, 69 y la 21, estos resultados son similares a las puntuaciones de la medida anterior, esto es sustentado por el alta correlación de dichas medidas (el coeficiente de correlación entre ambas fue de 0.94)

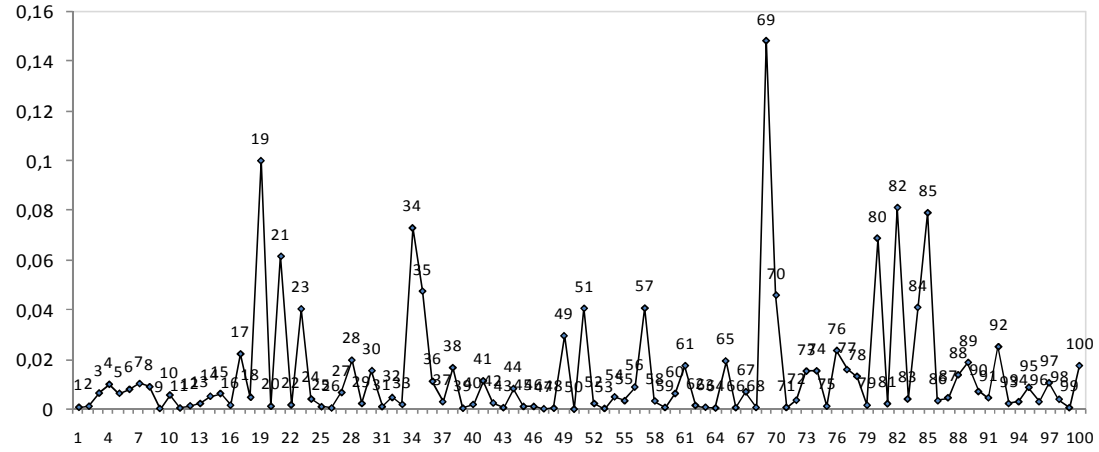


Las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación, según fórmula (38), se muestra en el GRÁFICO Nº 21, las observaciones potencialmente influyentes son la 69, 19, 82, 85 y la 80, ninguna de estas observaciones fueron mal clasificadas según la la función discriminante lineal de Fisher.



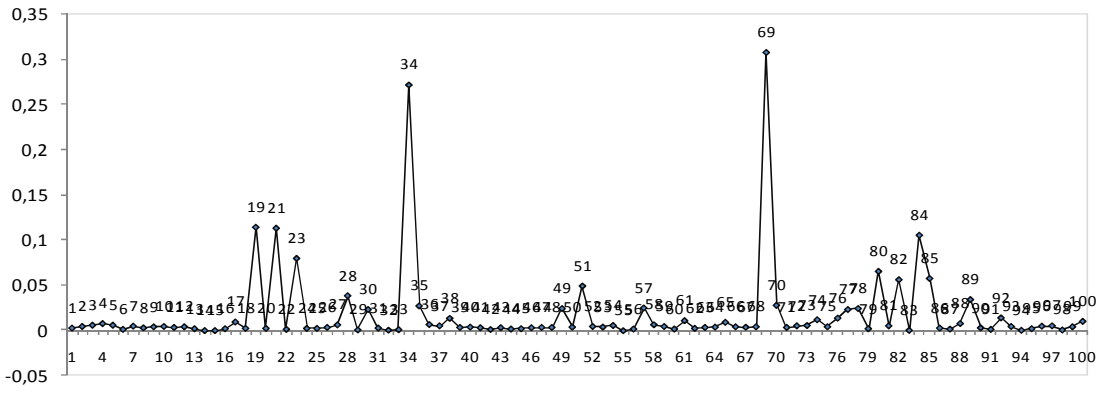
La representación gráfica de las puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de Taylor, según fórmula (42), se muestra en el GRÁFICO Nº 22, las observaciones que puedes considerarse como influyentes son la 69, 19, 82, 85, 84 y la 30.

GRÁFICO Nº 22
Puntuaciones de la medida de influencia alternativa para las probabilidades de Mala Clasificación, mediante aproximación de Taylor



Las puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher según, fórmula (43), se muestran en el GRÁFICO Nº 22, las observaciones potencialmente influyentes son 69, 34, 19, 21 y la 84.

GRÁFICO Nº 23
Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher.



En la TABLA N° 6, se muestra las puntuaciones de las medidas de influencia adicionales para los dos grupos, las observaciones potencialmente influyentes son, según la primera medida la 34 del primer grupo y la 84 del segundo grupo y según la segunda medida la 19 y la 34 del primer grupo y la 69 del segundo grupo.

TABLA N° 6 Medidas de influencia adiconales							
Primera Medida				Segunda Medida			
Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1	Obs.	Pun. Grupo G_1
34	65.112	84	46.761	19	9.516	69	15.978
21	43.843	69	39.795	34	8.284	85	12.824
23	38.965	89	32.762	49	7.172	82	12.787
19	32.554	80	30.963	21	6.626	68	10.194
28	30.343	78	30.057	35	6.498	92	9.666
35	20.808	77	29.570	11	6.130	57	9.112
17	16.473	74	23.059	1	5.298	73	8.923
30	15.387	61	20.610	38	5.222	65	8.355
49	13.017	51	19.215	23	5.215	51	8.144
29	11.884	70	17.723	13	5.114	80	7.296

1/ solo se presentan las 10 mayores puntuaciones de cada grupo

Teniendo en cuenta las puntuaciones de las distintas medidas de influencia para este conjunto de datos, podemos considerar observaciones influyentes a la 34 a la 19 y a la 69, pues la omisión de ellas, altera la estimación de los siguientes parámetros:

- La estimación de la distancia de Mahalanobis de 14.2889 a 15.4927; a 14.7716 y a 15.0270 respectivamente.
- Las estimaciones de las probabilidades de mala clasificación de 0.0297 a 0.0245; a 0.0273 y a 0.0263 respectivamente.
- Altera la tasa de error aparente de 3.0 a 2.02; 3.03 y a 3.04 respectivamente.

CONCLUSIONES

- Se demostró que existe una relación de proporcionalidad entre los coeficientes de la función lineal discriminante de Fisher en dos grupos y los coeficientes de la función de regresión lineal múltiple. Dicha relación no solo nos ha permitido plantearnos si es posible usar las medidas de influencia utilizadas en el análisis de regresión lineal múltiple, si no también permitió tomar como referencia, la metodología utilizada para derivar las medidas de influencia en el análisis de regresión y emularlo en el análisis discriminante lineal.

- Se determinó el efecto de una observación discordante en las estimaciones de los parámetros multivariantes, lo que ha permitido cuantificar la magnitud del efecto de dicha observación y hacer el tratamiento adecuado.

- Simulando un conjunto de datos con observaciones discordantes y luego identificándolas como tales, con las medidas de influencia estudiadas, se ha comprobado que realmente con dichas medidas estudiadas se detectan observaciones que difieren del resto. La simulación se convierte en una herramienta valiosa, pues en el presente trabajo ha permitido evaluar el potencial de las medidas de influencia propuestas.

- En los casos reales estudiados, usando las medidas de influencia, algunas de las observaciones que fueron mal clasificadas fueron identificadas como observaciones influyentes, mientras que, algunas otras observaciones que habían sido bien clasificadas, fueron detectadas como observaciones influyentes. Los resultados fueron:

    - Primera aplicación: Teniendo en cuenta las puntuaciones de las distintas medidas de influencia para este conjunto de datos, son identificadas como influyentes las observaciones 34,19 y 69, pues la omisión de ellas altera:

      Las estimaciones de las distancias de Mahalanobis pasan de 14.2889 a 15.4927, a 14.7716 y a 15.0270 respectivamente.

      Las estimaciones de las probabilidades de mala clasificación pasan de 0.0297 a 0.0245, a 0.0273 y a 0.0263 respectivamente.

      Las estimaciones de la tasa de error aparente pasan de 3.0 a 2.02; a 3.03 y a 3.04 respectivamente.

Segunda aplicación: Los resultados de las distintas medidas de influencia, indican que las observaciones influyentes son la 90, la 100 y la 64 pues la omisión de ellas, altera :

La estimación de la distancia de Mahalanobis, pasa de 4.6508 a 5.1117, a 5.0498 y a 4.9960 respectivamente.

Las estimaciones de las probabilidades de mala clasificación pasan de 0.1405 a 0.1298; a 0.1306 y a 0.1319 respectivamente.

Las estimaciones de la tasa de error aparente pasan de 8.0909 a 8.0808; y a 9.0909.

Tercera aplicación: Teniendo en cuenta las puntuaciones de las distintas medidas de influencia para este conjunto de datos, son influyentes las observaciones 34, 19 y 69, pues la omisión de ellas altera:

La estimación de la distancia de Mahalanobis pasa de 14.2889 a 15.4927; a 14.7716 y a 15.0270 respectivamente.

Las estimaciones de las probabilidades de mala clasificación pasa de 0.0297 a 0.0245; a 0.0273 y a 0.0263 respectivamente.

La estimación de la tasa de error aparente pasa de 3.0 a 2.02; 3.03 y a 3.04 respectivamente.

REFERENCIAS BIBLIOGRÁFICAS

1. Anscombe, F. J., and TUKEY, J. W. (1963), "*The Examination and Analysis of Residuals*" Technometrics, 5, 141-159.
2. Atkinson, et. al (2004). *A exploring multivariate data with the forward search*. New York: Springer.
3. Anderson T.W. (1984), *An introduction to Multivariate Statistics Analysis*, 2 ed., Wiley e Sons, Inc, N.Y.
4. Beckman, R.J. y Cook, R.D. (1983). *Outlier…….s*. American Statistical Association and American Society for Quality: Technometrics, Vol. 25, pp 119-149
5. Belsley, D. Kuh, E., Welsch,R. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley &Sons.
6. Campbell, N.A. (1978). *The Influence Function as an Aid in Outlier Detection in Discriminant Analysis*. Applied. Statistics, 27, 251-258.
7. Castaño, E. (1988). *Observaciones Influénciales*. Revista Colombiana de Estadística. Universidad de Antioquia–Colombia. N° 17-18.
8. Cook, R.D.; Weisberg, s. (1982) *Residual and Influence in Regresión*. Chapman & Hall.
9. Devlin, S.J.; Gananadesikan, R.; Kettenring, J.R. (1975). *Robust Estimation and Outlier Detection with correlation coefficients*. Biometrika, 62, 531 – 545.
10. Enguix, A. (2001). *Análisis de Influencia en Componentes Principales*. Universidad de Sevilla-España. Facultad de Matemáticas.
11. Furtado, D. (2008). *Estatística multivariada*. Primera edición. Lavras-Brasil:Ed. UFLA.
12. Fung, W.K. (1992). Diagnostics *in Linear Discriminant analysis*. Statistics and Probability Letters, 13, 279–285.
13. Fung, W.K. (1995). *Some Diagnostic Measures in Discriminant Analysis*. Journal of the American Statistical Association, 90, 952-956.
14. Gómez, D; Solano, O; Albán J; Vásquez C; Adriazola Y; Quinteros Y. (2008). *Determinación de patrones de variación morfológica del género Minthostachys en Unchus y Cajatambo, mediante métodos estadísticos multivariantes de reducción de datos*. Facultad de Matemática, UNMSM. Pesquimat. Vol.XI-N 1Pgs.45-56. Lima-Peru.
15. Hair,J. et. Al (1999). *Análisis multivariante*. Madrid: Prentice Hall.
16. Hampel, F.R. (1974). *Influence Curve and Its Role in Robust Estimation*. Journal of the American Statistical Association, 69, 383-393.

17. Johnson, D. (2000). *Métodos multivariados aplicados al análisis de datos.* Méjico: International Thomson Editores.
18. Jhonson, R.A. (1982). *Applied Multivariante Statistical Analysis*. Glenwood. Prentice Hall.
19. Lachenbruch, P. A. (1975) *Discriminant Analysis*. New York: Hafner.
20. Lachenbruch , P.A.; Mickey, M.R. (1968). *Estimation of Error Rates in Discriminant Analysis. Washington*. Thechnometric. V.10; N 1, p 1-11.
21. Manly,Bryan.(2005), *Multivariate statistical methods*. A primer. Tercera edición. New York. Ed. Chapman & Hall/CRC.
22. Mardia,K. et al (1979) . *Multivariate analysis. London*: Academic Press.
23. Maronna, R., et al. (2006). *Robust statistics*: Theory and methods. John Wiley & Sons Inc.
24. Morillas, A. y Díaz B. (2007). *El Problema de los Outliers Multivariantes en el Análisis de Sectores Cave y Cluster Industrial.* Universidad de Málaga.
25. Muños, J.M; Moreno, J.L; Gómez, T; Enguix, A. (2001). *Sesgo Condicionado en el Análisis de Influencia una Revisión*. Facultad de Matemática, Universidad de Sevilla. Questhó. 25, 263-284.
26. Peña, D. (2002). *Análisis de datos Multivariantes.* McGraw-Hill/Interamericana de España.

APÉNDICE

A. CÁLCULOS ADICIONALES

A1 Caso general de la igualdad de Sherman-Morrison

Considerase la matriz $X^T X$ de orden pxp y sea x^T el i-ésimo reglón de X obsérvese que $X^T X - xx^T$ es la matriz $X^T X$ con su i-ésimo reglón eliminado, sea A y B constante en R, entonces se cumple lo siguiente:

$$[X^T X - xx^T]^{-1} = A^{-1}\left[(X^T X)^{-1} + \frac{B(X^T X)^{-1} xx^T (X^T X)^{-1}}{A - Bx^T (X^T X)^{-1} x}\right]$$

En efecto: esta situación se resuelve, multiplicando ambos miembros por la cantidad $[AX^T X - Bxx^T]$

$$[X^T X - xx^T]^{-1}[AX^T X - Bxx^T] = A^{-1}\left[(X^T X)^{-1} + \frac{B(X^T X)^{-1} xx^T (X^T X)^{-1}}{A - Bx^T (X^T X)^{-1} x}\right][AX^T X - Bxx^T]$$

$$= A^{-1}\left[A(X^T X)^{-1}(X^T X) + \frac{BA(X^T X)^{-1} xx^T (X^T X)^{-1}(X^T X)}{A - Bx^T (X^T X)^{-1} x} - (X^T X)^{-1} Bxx^T - \frac{B^2 (X^T X)^{-1} xx^T (X^T X)^{-1} xx^T}{A - Bx^T (X^T X)^{-1} x}\right]$$

$$= A^{-1}\left[A + \frac{BA(X^T X)^{-1} xx^T}{A - Bx^T (X^T X)^{-1} x} - (X^T X)^{-1} Bxx^T - \frac{B^2 (X^T X)^{-1} x\left[x^T (X^T X)^{-1} x\right]x^T}{A - Bx^T (X^T X)^{-1} x}\right]$$

$$= A^{-1}\left[A + \frac{BA(X^T X)^{-1} xx^T - BA(X^T X)^{-1} xx^T + B^2 (X^T X)^{-1} xx^T \left[x^T (X^T X)^{-1} x\right] - B^2 (X^T X)^{-1} xx^T \left[x^T (X^T X)^{-1} x\right]}{A - Bx^T (X^T X)^{-1} x}\right]$$

Eliminando los términos semejantes con distintos signos de obtiene:

$$[X^T X - xx^T]^{-1}[AX^T X - Bxx^T] = I$$

A2

Si $\tilde{\tilde{\Sigma}} \rightarrow \ (1 - w_1\varepsilon)\Sigma_{F_1} + \varepsilon.w_1.\vec{Z}.\vec{Z}^T$ entonces

$$\tilde{\tilde{\Sigma}}^{-1} \rightarrow (1-\varepsilon w_1)\left(.\Sigma^{-1} - \frac{\varepsilon w_1.\Sigma^{-1}\vec{Z}.\vec{Z}^T.\Sigma^{-1}}{1-\varepsilon w_1 + \varepsilon w_1 \vec{Z}^T.\Sigma^{-1}Z} \right)$$

En efecto,

Esta igualdad, se obtiene haciendo $A = (1-\varepsilon w_1)$; $B = \varepsilon w_1$ $X^T X = \Sigma$ y $x = \vec{Z}$ en el caso general de la igualdad de Sherman-Morrison demostrado en **A1.**

$$\tilde{\tilde{\Sigma}}^{-1} \rightarrow (1-\varepsilon w_1)\left(.\Sigma^{-1} - \frac{\varepsilon w_1.\Sigma^{-1}\vec{Z}.\vec{Z}^T.\Sigma^{-1}}{1-\varepsilon w_1 + \varepsilon w_1 \vec{Z}^T.\Sigma^{-1}Z} \right) = (1-\varepsilon w_1) - \varepsilon w_1.\Sigma^{-1}\vec{Z}.\vec{Z}^T.\Sigma^{-1}$$

Sea $I\left(\vec{X};\Delta^2\right) = w_1\Delta^2 + 2\psi - w_1\psi^2$, donde $\Delta^2, \psi, w_1 \in R$, entonces

$$I_{\max}(X,\Delta^2) = w_1.\Delta^2 + w_1^{-1}$$

En efecto: Como $\Delta^2, \psi, w_1 \in R$ se tiene lo siguiente:

Agrupando, convenientemente los términos de la expresión $I\left(\vec{X};\Delta^2\right)$, además agregando y quitando el término w_1^{-2}

$$I\left(\vec{X};\Delta^2\right) = w_1\Delta^2 - w_1(\psi^2 - 2\psi w_1^{-1} - w_1^{-2} + w_1^{-2})$$

$$I\left(\vec{X};\Delta^2\right) = \underbrace{w_1\Delta^2 + w_1^{-1}}_{F} - \underbrace{w_1(\psi - w_1^{-1})^2}_{G} \quad \text{.....(A)}$$

De la expresión obtenida en (A), se puede hacer el siguiente análisis.

La expresión "F" es definida positiva debido a lo siguiente:

- Δ^2 Por ser definida como distancia es definida positiva.
- w_1 es el peso del primer grupo en la formación de la matriz de varianzas y covarianzas muestral S_u, además como $w_1 > 0 \Rightarrow w_1^{-1} > 0$

Con estas consideraciones, podemos afirmar que la expresión "F" es definida positiva.

Entonces la expresión $I\left(\vec{X};\Delta^2\right)$ definida en (A) mediante "F" y "G" alcanzará su máximo valor siempre y cuando "G" sea mínimo, y como "G" es una expresión cuadrática definida como $(\psi - w_1^{-1})^2$, será mínimo cuando sea igual a cero y el valor que hace a "G" mínimo es $\psi = w_1^{-1}$, finalmente el valor máximo de $I\left(\vec{X};\Delta^2\right)$ será:

$$I_{\max}(X,\Delta^2) = w_1.\Delta^2 + w_1^{-1}$$

A3

Demostrar que $\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)} - \vec{\overline{X}}\right)^T = 0$

En efecto:

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)} - \vec{\overline{X}}\right)^T = \sum_{i=1}^{n_1}\left(\vec{X}_i^{(1)} - \vec{\overline{X}}^{(1)}\right).\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}\right)^T + \sum_{i=1}^{n_2}\left(\vec{X}_i^{(2)} - \vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(2)} - \vec{\overline{X}}\right)^T$$

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)} - \vec{\overline{X}}\right)^T = \left(\sum_{i=1}^{n_1}\vec{X}_i^{(1)} - n_1.\vec{\overline{X}}^{(1)}\right).\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}\right)^T + \left(\sum_{i=1}^{n_2}\vec{X}_i^{(2)} - n_2.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(2)} - \vec{\overline{X}}\right)^T$$

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right).\left(\vec{\overline{X}}^{(k)} - \vec{\overline{X}}\right)^T = \left(n_1.\vec{\overline{X}}^{(1)} - n_1.\vec{\overline{X}}^{(1)}\right).\left(\vec{\overline{X}}^{(1)} - \vec{\overline{X}}\right)^T + \left(n_2.\vec{\overline{X}}^{(2)} - n_2.\vec{\overline{X}}^{(2)}\right).\left(\vec{\overline{X}}^{(2)} - \vec{\overline{X}}\right)^T = 0$$

Con lo que queda demostrado, del mismo modo se puede demostrar lo siguiente:

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\vec{\overline{X}}^{(k)} - \vec{\overline{X}}\right).\left(\vec{X}_i^{(k)} - \vec{\overline{X}}^{(k)}\right)^T = 0$$

B. DATOS SIMULADOS Y RESULTADOS IMPORTANTES

TABLA N° B1
Datos simulado por el método de Monte Carlo

Grupo G_1					Grupo G_2				
Obs.	X_1	X_2	X_3	X_4	Obs.	X_1	X_2	X_3	X_4
1	1,58	1,65	6,88	-1,42	26	0,43	5,57	-2,86	-0,68
2	-0,20	2,79	-0,68	0,18	27	0,68	2,95	2,46	-1,74
3	0,72	2,29	0,83	-0,93	28	-1,63	-0,01	6,18	6,67
4	1,89	2,79	0,80	-4,16	29	1,39	-3,24	5,93	-1,77
5	2,68	3,56	3,67	-1,98	30	1,10	3,85	1,38	-1,67
6	-1,26	-0,54	3,97	5,51	31	0,46	1,75	3,15	2,05
7	1,32	6,34	-3,24	-3,36	32	2,05	5,45	0,53	-2,35
8	-0,73	0,23	3,35	7,07	33	-0,62	0,34	5,97	5,08
9	0,22	1,55	1,89	-1,24	34	-1,08	2,56	0,01	1,54
10	0,52	4,08	0,71	-0,86	35	0,71	1,95	4,72	0,86
11	-2,52	-1,72	-0,35	2,58	36	-1,41	2,57	2,72	5,88
12	-2,04	-0,02	3,62	5,88	37	-1,74	-1,14	2,52	4,07
13	-1,05	-5,13	3,20	1,37	38	-0,12	-1,83	6,03	0,67
14	-0,42	0,54	0,18	1,62	39	-0,69	-1,14	5,40	2,43
15	2,31	7,00	-1,83	-3,71	40	-1,25	-2,05	2,92	3,98
16	1,81	1,91	2,24	1,96	41	0,58	2,84	-0,03	-1,57
17	-1,26	-0,54	3,97	5,51	31	0,46	1,75	3,15	2,05
18	1,32	6,34	-3,24	-3,36	32	2,05	5,45	0,53	-2,35
19	-0,73	0,23	3,35	7,07	33	-0,62	0,34	5,97	5,08
20	0,22	1,55	1,89	-1,24	34	-1,08	2,56	0,01	1,54
21	0,52	4,08	0,71	-0,86	35	0,71	1,95	4,72	0,86
22	2,68	3,56	3,67	-1,98	30	1,10	3,85	1,38	-1,67
23	-1,26	-0,54	3,97	5,51	31	0,46	1,75	3,15	2,05
24	1,32	6,34	-3,24	-3,36	32	2,05	5,45	0,53	-2,35
25	-0,73	0,23	3,35	7,07	33	-0,62	0,34	5,97	5,08
26	0,22	1,55	1,89	-1,24	34	-1,08	2,56	0,01	1,54
27	0,52	4,08	0,71	-0,86	35	0,71	1,95	4,72	0,86
28	-2,52	-1,72	-0,35	2,58	36	-1,41	2,57	2,72	5,88
29	-2,04	-0,02	3,62	5,88	37	-1,74	-1,14	2,52	4,07
30	-1,05	-5,13	3,20	1,37	38	-0,12	-1,83	6,03	0,67
31	-0,42	0,54	0,18	1,62	39	-0,69	-1,14	5,40	2,43
32	2,31	7,00	-1,83	-3,71	40	-1,25	-2,05	2,92	3,98
33	1,81	1,91	2,24	1,96	41	0,58	2,84	-0,03	-1,57
34	-1,26	-0,54	3,97	5,51	31	0,46	1,75	3,15	2,05
35	1,32	6,34	-3,24	-3,36	32	2,05	5,45	0,53	-2,35
36	-0,73	0,23	3,35	7,07	33	-0,62	0,34	5,97	5,08
37	0,22	1,55	1,89	-1,24	34	-1,08	2,56	0,01	1,54
38	0,52	4,08	0,71	-0,86	35	0,71	1,95	4,72	0,86
39	-2,52	-1,72	-0,35	2,58	36	-1,41	2,57	2,72	5,88
40	-2,04	-0,02	3,62	5,88	37	-1,74	-1,14	2,52	4,07
41	-1,05	-5,13	3,20	1,37	38	-0,12	-1,83	6,03	0,67
42	-0,42	0,54	0,18	1,62	39	-0,69	-1,14	5,40	2,43
43	2,31	7,00	-1,83	-3,71	40	-1,25	-2,05	2,92	3,98
44	1,81	1,91	2,24	1,96	41	0,58	2,84	-0,03	-1,57
45	-0,97	-1,28	1,46	2,51	42	-1,26	1,19	0,36	2,54
46	-0,36	-1,33	1,53	0,86	43	0,71	-1,18	1,50	-1,62
47	-2,51	0,64	1,37	7,04	44	0,73	1,33	2,03	2,17
48	-1,07	1,87	1,16	1,46	45	-0,34	1,05	5,86	5,29
49	2,88	2,38	5,12	-2,07	46	0,32	4,42	-1,27	-1,09
50	-3,04	-6,26	6,07	7,5	50	-1,32	2,62	-1,91	3,73

TABLA N° B2

Diferencia de cada observación respecto al vector de medias para los datos de Gorriones

Grupo G_1 : Gorriones que no sobrevivieron a la tormenta						Grupo G1 : Gorriones que sobrevivieron a la tormenta					
Obs.	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**Obs.**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**
1	-1.38	4.00	0.17	-	-0.31	22	-3.43	-1.57	-0.08	-0.45	-0.14
2	-3.38	-1.00	-1.03	-0.60	-1.21	23	-2.43	-1.57	0.02	-0.25	-0.24
3	-4.38	-1.00	-0.43	-0.10	-0.21	24	1.57	0.43	1.12	0.35	0.86
4	-4.38	-5.00	-0.53	-0.80	-0.61	25	-6.43	-9.57	-1.18	-1.25	-1.04
5	-2.38	2.00	0.07	0.10	-0.51	26	1.57	8.43	0.22	0.35	1.66
6	5.62	6.00	0.57	0.50	0.09	27	-3.43	-4.57	-0.48	0.05	-0.84
7	-0.38	-3.00	-0.53	-0.10	-0.61	28	-1.43	3.43	0.72	1.05	0.56
8	-2.38	-2.00	1.37	0.10	0.39	29	6.57	3.43	1.62	1.35	1.86
9	6.62	7.00	1.27	0.60	0.29	30	-5.43	-10.57	-1.38	-1.15	-1.04
10	0.62	-3.00	-0.43	0.30	1.19	31	3.57	-2.57	-1.18	-0.45	2.26
11	0.62	-1.00	-0.13	0.10	1.19	32	3.57	1.43	0.12	0.35	0.46
12	2.62	3.00	-0.33	0.10	-0.31	33	0.57	3.43	0.32	0.05	0.86
13	3.62	5.00	0.87	0.80	0.99	34	0.57	5.43	-0.58	-0.35	-1.84
14	-0.38	4.00	0.57	0.60	-0.81	35	-3.43	1.43	-0.58	0.05	0.46
15	-0.38	-6.00	0.07	-0.40	-1.01	36	3.57	10.43	0.42	0.65	1.36
16	-1.38	-4.00	-0.53	-0.50	-0.51	37	-6.43	-11.57	-1.08	-1.15	-2.24
17	0.62	3.00	-0.03	-	0.79	38	0.57	0.43	-0.68	-0.25	-0.34
18	-4.38	-3.00	-0.93	-0.30	0.09	39	-3.43	-3.57	-0.28	-0.55	-1.54
19	-2.38	-5.00	-1.13	-	-0.71	40	4.57	7.43	1.92	1.05	1.96
20	5.62	5.00	1.07	0.10	1.09	41	4.57	0.43	-0.48	-0.35	-0.14
21	1.62	-5.00	0.07	-0.50	0.69	42	-2.43	-4.57	0.22	-0.25	-0.54
						43	0.57	-3.57	0.02	-0.05	-0.54
						44	2.57	3.43	0.62	0.65	-0.04
						45	-3.43	-6.57	-0.78	-0.75	-1.24
						46	3.57	5.43	0.42	0.65	-0.44
						47	-5.43	-4.57	-0.88	0.15	-0.44
						48	3.57	3.43	1.02	0.05	0.26
						49	5.57	6.43	0.82	0.35	0.06

TABLA N° B3

Diferencia de cada observación respecto al vector de medias para los datos de Minthostachys (Muña)

Grupo G_1 :Pubescencia Abundante				Grupo G_2 : Pubescencia escasa			
Obs.	Peciolo	L hoja	A hoja	Obs.	Peciolo	L hoja	A hoja
1	0.02	0.36	0.06	52	0.30	0.23	0.33
2	-0.18	-0.24	-0.54	53	-0.70	-0.67	-0.87
3	0.02	0.06	-0.24	54	-0.20	0.13	0.13
4	0.02	-0.34	0.06	55	0.20	-0.17	-0.17
5	0.02	0.46	0.26	56	0.90	0.33	0.33
6	0.02	0.06	-0.34	57	-0.00	-0.07	-0.37
7	-0.18	0.46	-0.24	58	0.40	-0.17	-0.17
8	0.02	-0.14	-0.14	59	0.10	0.93	0.53
9	0.02	-0.44	-0.24	60	-0.00	0.73	-0.27
10	0.22	-0.14	0.26	61	0.20	-0.67	-0.27
11	0.22	0.26	0.26	62	-0.60	-0.87	-1.07
12	-0.08	-0.04	0.46	63	0.10	-0.37	-0.07
13	-0.08	0.66	-0.44	64	1.20	0.43	0.63
14	-0.08	0.26	-0.39	65	0.90	0.93	0.43
15	0.02	0.06	-0.44	66	0.50	0.93	0.33
16	0.12	-0.24	0.16	67	0.50	0.93	1.13
17	0.22	1.16	0.16	68	0.30	0.83	-0.07
18	0.22	1.46	0.16	69	-0.20	-0.57	-0.37
19	-0.08	-0.24	0.06	70	-0.10	0.83	0.33
20	0.12	-0.44	-0.54	71	-0.30	-0.67	-0.17
21	0.12	-1.14	0.06	72	-0.20	-0.17	-0.07
22	0.22	0.46	0.96	73	-0.00	0.73	0.43
23	-0.08	-0.34	0.06	74	-0.60	-0.57	-0.57
24	0.12	0.16	0.36	75	-0.50	-0.77	-0.27
25	0.02	0.66	0.06	76	0.10	-0.07	0.33
26	0.12	0.06	0.06	77	-0.10	0.53	0.63
27	-0.18	-0.54	-0.24	78	-0.10	0.83	0.43
28	0.12	-0.84	-0.29	79	-0.00	0.93	0.53
29	0.12	0.56	0.31	80	0.20	1.13	0.73
30	-0.08	-1.34	-0.24	81	-0.00	0.43	0.43
31	0.12	0.56	0.06	82	0.90	0.83	0.23
32	-0.08	-1.24	-0.64	83	-0.30	-0.37	-0.37
33	0.02	-0.44	-0.24	84	-0.10	0.53	0.33
34	-0.13	0.36	-0.04	85	-0.40	-0.87	-0.77
35	0.02	0.36	0.16	86	-0.00	0.43	0.23
36	0.02	-0.54	-0.44	87	0.40	0.33	0.23
37	-0.08	0.66	0.36	88	-0.40	-1.07	-0.67
38	0.02	0.06	0.06	89	-0.20	-0.67	-0.47
39	-0.18	0.06	0.16	90	-1.00	-1.77	-1.07
40	-0.08	0.66	0.26	91	-0.70	-1.37	-0.77
41	0.02	-0.04	0.46	92	-0.00	0.93	0.33
42	-0.18	-0.14	-0.24	93	-0.10	-0.57	0.03
43	-0.28	-0.44	-0.44	94	1.00	0.43	0.73
44	0.12	0.46	-0.14	95	0.30	-0.57	0.13
45	-0.08	-0.24	-0.24	96	-0.10	0.33	0.23
46	0.02	-0.24	0.96	97	0.50	0.73	0.23
47	-0.08	-0.34	0.36	98	-0.70	-1.37	-0.87
48	-0.08	-0.64	0.26	99	-0.40	-0.37	0.23
49	-0.18	0.26	0.26	100	-0.90	-1.47	-0.97
50	-0.18	0.16	-0.34				
51	-0.08	0.16	-0.24				

TABLA N° B4 Diferencia de cada observación respecto al vector de medias para los datos de iris									
Grupo G_1 : Versicolor					Grupo G_2 : Virginica				
Obs.	L Sepalo	A Sepalo	L petalo	A petalo	Obs.	L Sepalo	A Sepalo	L petalo	A petalo
1	1.06	0.43	0.44	0.07	26	-0.29	0.33	0.45	0.47
2	0.46	0.43	0.24	0.17	27	-0.79	-0.27	-0.45	-0.13
3	0.96	0.33	0.64	0.17	28	0.51	0.03	0.35	0.07
4	-0.44	-0.47	-0.26	-0.03	29	-0.29	-0.07	0.05	-0.23
5	0.56	0.03	0.34	0.17	30	-0.09	0.03	0.25	0.17
6	-0.24	0.03	0.24	-0.03	31	1.01	0.03	1.05	0.07
7	0.36	0.53	0.44	0.27	32	-1.69	-0.47	-1.05	-0.33
8	-1.04	-0.37	-0.96	-0.33	33	0.71	-0.07	0.75	-0.23
9	0.66	0.13	0.34	-0.03	34	0.11	-0.47	0.25	-0.23
10	-0.74	-0.07	-0.36	0.07	35	0.61	0.63	0.55	0.47
11	-0.94	-0.77	-0.76	-0.33	36	-0.09	0.23	-0.45	-0.03
12	-0.04	0.23	-0.06	0.17	37	-0.19	-0.27	-0.25	-0.13
13	0.06	-0.57	-0.26	-0.33	38	0.21	0.03	-0.05	0.07
14	0.16	0.13	0.44	0.07	39	-0.89	-0.47	-0.55	-0.03
15	-0.34	0.13	-0.66	-0.03	40	-0.79	-0.17	-0.45	0.37
16	0.76	0.33	0.14	0.07	41	-0.19	0.23	-0.25	0.27
17	-0.34	0.23	0.24	0.17	31	-0.09	0.03	-0.05	-0.23
18	-0.14	-0.07	-0.16	-0.33	32	1.11	0.83	1.15	0.17
19	0.26	-0.57	0.24	0.17	33	1.11	-0.37	1.35	0.27
20	-0.34	-0.27	-0.36	-0.23	34	-0.59	-0.77	-0.55	-0.53
21	-0.04	0.43	0.54	0.47	35	0.31	0.23	0.15	0.27
22	0.16	0.03	-0.26	-0.03	30	-0.99	-0.17	-0.65	-0.03
23	0.36	-0.27	0.64	0.17	31	1.11	-0.17	1.15	-0.03
24	0.16	0.03	0.44	-0.13	32	-0.29	-0.27	-0.65	-0.23
25	0.46	0.13	0.04	-0.03	33	0.11	0.33	0.15	0.07
26	0.66	0.23	0.14	0.07	34	0.61	0.23	0.45	-0.23
27	0.86	0.03	0.54	0.07	35	-0.39	-0.17	-0.75	-0.23
28	0.76	0.23	0.74	0.37	36	-0.49	0.03	-0.65	-0.23
29	0.06	0.13	0.24	0.17	37	-0.19	-0.17	0.05	0.07
30	-0.24	-0.17	-0.76	-0.33	38	0.61	0.03	0.25	-0.43
31	-0.44	-0.37	-0.46	-0.23	39	0.81	-0.17	0.55	-0.13
32	-0.44	-0.37	-0.56	-0.33	40	1.31	0.83	0.85	-0.03
33	-0.14	-0.07	-0.36	-0.13	41	-0.19	-0.17	0.05	0.17
34	0.06	-0.07	0.84	0.27	31	-0.29	-0.17	-0.45	-0.53
35	-0.54	0.23	0.24	0.17	32	-0.49	-0.37	0.05	-0.63
36	0.06	0.63	0.24	0.27	33	1.11	0.03	0.55	0.27
37	0.76	0.33	0.44	0.17	34	-0.29	0.43	0.05	0.37
38	0.36	-0.47	0.14	-0.03	35	-0.19	0.13	-0.05	-0.23
39	-0.34	0.23	-0.16	-0.03	36	-0.59	0.03	-0.75	-0.23
40	-0.44	-0.27	-0.26	-0.03	37	0.31	0.13	-0.15	0.07
41	-0.44	-0.17	0.14	-0.13	38	0.11	0.13	0.05	0.37
42	0.16	0.23	0.34	0.07	39	0.31	0.13	-0.45	0.27
43	-0.14	-0.17	-0.26	-0.13	40	-0.79	-0.27	-0.45	-0.13
44	-0.94	-0.47	-0.96	-0.33	41	0.21	0.23	0.35	0.27
45	-0.34	-0.07	-0.06	-0.03	42	0.11	0.33	0.15	0.47
46	-0.24	0.23	-0.06	-0.13	43	0.11	0.03	-0.35	0.27
47	-0.24	0.13	-0.06	-0.03	44	-0.29	-0.47	-0.55	-0.13
48	0.26	0.13	0.04	-0.03	45	-0.09	0.03	-0.35	-0.03
49	-0.84	-0.27	-1.26	-0.23	46	-0.39	0.43	-0.15	0.27
50	-0.24	0.03	-0.16	-0.03	50	-0.69	0.03	-0.45	-0.23

C. PUNTUACIONES DE LAS MEDIDAS DE INFLUENCIA

CONJUNTO DE DATOS SIMULADOS

TABLA N° B5

Puntuaciones de la medida de influencia para la distancia de Mahalanobis

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
10	83,72	43	8,77	58	3,86	80	1,09	27	0,25
16	50,93	84	8,12	6	3,78	59	1,09	86	0,25
98	31,65	31	7,67	91	3,77	20	1,08	52	0,22
65	25,27	15	6,99	50	3,75	100	1,05	61	0,22
8	24,01	32	6,80	4	3,73	97	1,03	82	0,21
79	22,12	67	6,58	90	3,64	94	0,94	88	0,19
44	21,91	95	6,41	3	3,53	11	0,88	83	0,19
93	18,03	14	6,31	23	3,50	30	0,87	42	0,16
21	17,94	51	6,27	75	3,49	2	0,84	47	0,14
49	16,22	24	5,62	74	3,05	25	0,78	1	0,12
92	16,18	87	5,48	26	2,60	69	0,70	56	0,10
89	15,79	66	5,37	68	2,53	99	0,62	48	0,07
5	15,52	73	5,21	22	2,47	53	0,57	62	0,06
60	15,44	18	5,13	33	2,34	54	0,56	81	0,06
76	12,70	64	4,80	7	2,04	77	0,56	96	0,03
70	11,82	45	4,71	63	1,90	19	0,49	28	0,03
85	11,27	40	4,71	71	1,76	36	0,49	38	0,02
57	11,05	17	4,26	35	1,46	34	0,45	39	0,01
29	10,25	13	4,20	78	1,27	46	0,34	9	0,01
55	8,78	72	4,09	41	1,16	37	0,26	12	0,01

TABLA N° B6

Puntuaciones de la medida de influencia para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
10	-0,34	39	-0,04	30	-0,02	4	0,02	84	0,05
55	-0,14	38	-0,04	94	-0,01	6	0,02	43	0,05
95	-0,13	28	-0,04	100	-0,01	90	0,02	29	0,06
87	-0,12	47	-0,03	97	-0,01	50	0,02	85	0,06
72	-0,11	1	-0,03	80	-0,01	58	0,02	57	0,06
20	-0,08	42	-0,03	41	-0,01	17	0,02	70	0,06
59	-0,08	88	-0,03	78	-0,01	13	0,02	76	0,07
11	-0,07	83	-0,03	35	-0,01	40	0,02	60	0,08
34	-0,07	61	-0,03	71	-0,00	45	0,02	5	0,08
27	-0,06	37	-0,03	63	-0,00	64	0,02	89	0,08
86	-0,06	46	-0,03	7	0,00	18	0,03	49	0,08
52	-0,06	36	-0,02	33	0,00	73	0,03	92	0,08
82	-0,06	19	-0,02	68	0,01	66	0,03	21	0,09
56	-0,05	54	-0,02	22	0,01	24	0,03	93	0,09
81	-0,05	77	-0,02	26	0,01	14	0,03	44	0,10
62	-0,05	53	-0,02	74	0,01	51	0,04	79	0,10
48	-0,05	99	-0,02	3	0,01	67	0,04	8	0,11
96	-0,05	69	-0,02	75	0,02	32	0,04	65	0,11
9	**-0,05**	25	-0,02	23	0,02	15	0,04	98	0,13
12	-0,05	2	-0,02	91	0,02	31	0,04	16	0,18

TABLA N° B7

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
10	0,0158	66	0,0001	17	0,0000	42	0,0000	70	0,0000
8	0,0003	73	0,0001	18	0,0000	44	0,0000	71	0,0000
21	0,0003	76	0,0001	19	0,0000	45	0,0000	72	0,0000
79	0,0003	85	0,0001	20	0,0000	46	0,0000	74	0,0000
93	0,0003	87	0,0001	25	0,0000	47	0,0000	75	0,0000
98	0,0003	89	0,0001	26	0,0000	48	0,0000	77	0,0000
16	0,0002	90	0,0001	27	0,0000	51	0,0000	78	0,0000
29	0,0002	92	0,0001	28	0,0000	52	0,0000	80	0,0000
57	0,0002	95	0,0001	30	0,0000	53	0,0000	81	0,0000
5	0,0001	1	0,0000	31	0,0000	54	0,0000	82	0,0000
15	0,0001	2	0,0000	32	0,0000	56	0,0000	83	0,0000
22	0,0001	3	0,0000	33	0,0000	58	0,0000	84	0,0000
23	0,0001	4	0,0000	34	0,0000	59	0,0000	86	0,0000
24	0,0001	6	0,0000	35	0,0000	61	0,0000	88	0,0000
43	0,0001	7	0,0000	36	0,0000	62	0,0000	91	0,0000
49	0,0001	9	0,0000	37	0,0000	63	0,0000	94	0,0000
50	0,0001	11	0,0000	38	0,0000	64	0,0000	96	0,0000
55	0,0001	12	0,0000	39	0,0000	67	0,0000	97	0,0000
60	0,0001	13	0,0000	40	0,0000	68	0,0000	99	0,0000
65	0,0001	14	0,0000	41	0,0000	69	0,0000	100	0,0000

TABLA N° B8

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de taylor

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
10	0,0033	6	-	31	-	51	-	75	-
79	0,0002	7	-	32	-	52	-	76	-
5	0,0001	9	-	33	-	53	-	77	-
8	0,0001	11	-	34	-	54	-	78	-
16	0,0001	12	-	35	-	56	-	80	-
21	0,0001	13	-	36	-	58	-	81	-
29	0,0001	14	-	37	-	59	-	82	-
55	0,0001	15	-	38	-	61	-	83	-
57	0,0001	17	-	39	-	62	-	84	-
60	0,0001	18	-	40	-	63	-	85	-
65	0,0001	19	-	41	-	64	-	86	-
87	0,0001	20	-	42	-	66	-	88	-
92	0,0001	22	-	43	-	67	-	89	-
93	0,0001	23	-	44	-	68	-	90	-
95	0,0001	24	-	45	-	69	-	91	-
98	0,0001	25	-	46	-	70	-	94	-
1	-	26	-	47	-	71	-	96	-
2	-	27	-	48	-	72	-	97	-
3	**-**	28	-	49	-	73	-	99	-
4	-	30	-	50	-	74	-	100	-

TABLA N° B9 Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
10	11,8580	54	0,0118	62	0,0097	7	0,0049	72	0,0036
16	0,2674	77	0,0118	68	0,0097	67	0,0049	28	0,0035
98	0,0701	83	0,0118	81	0,0097	45	0,0048	38	0,0035
8	0,0594	61	0,0117	23	0,0096	26	0,0047	39	0,0035
79	0,0419	80	0,0117	56	0,0095	25	0,0046	6	0,0032
21	0,0348	94	0,0116	91	0,0095	43	0,0045	12	0,0032
93	0,0314	60	0,0115	74	0,0094	84	0,0045	9	0,0031
65	0,0289	78	0,0115	89	0,0093	19	0,0043	48	0,0030
44	0,0236	88	0,0115	82	0,0089	4	0,0041	32	0,0029
55	0,0196	92	0,0115	52	0,0088	33	0,0041	11	0,0027
29	0,0194	49	0,0114	86	0,0088	36	0,0041	27	0,0025
5	0,0179	75	0,0114	51	0,0079	46	0,0039	17	0,0024
57	0,0174	100	0,0114	22	0,0070	1	0,0038	40	0,0024
90	0,0140	63	0,0112	64	0,0069	2	0,0038	70	0,0024
95	0,0140	71	0,0107	85	0,0069	30	0,0038	34	0,0023
53	0,0125	73	0,0102	50	0,0062	37	0,0038	3	0,0022
66	0,0125	87	0,0102	59	0,0060	47	0,0038	18	0,0021
69	0,0123	76	0,0101	15	0,0056	35	0,0037	14	0,0019
97	0,0123	58	0,0100	24	0,0054	41	0,0037	20	0,0018
99	0,0119	96	0,0099	13	0,0051	42	0,0037	31	0,0012

PRIMER CONJUNTO DE DATOS

TABLA N° B10 Puntuaciones de la medida de influencia para la distancia de mahalanobis									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
28	5,33	29	2,91	2	2,46	46	2,17	14	1,51
47	5,28	42	2,74	4	2,36	33	2,16	18	1,39
20	4,57	16	2,73	25	2,36	36	2,15	19	1,37
27	3,95	23	2,65	30	2,34	13	2,12	3	1,26
9	3,94	40	2,64	26	2,33	11	2,10	34	1,21
35	3,79	44	2,60	39	2,31	32	1,95	48	1,21
21	3,69	17	2,59	7	2,27	38	1,76	31	1,18
6	3,61	22	2,58	1	2,21	5	1,65	49	1,10
12	3,16	37	2,48	45	2,20	10	1,59	41	0,73
15	2,92	24	2,48	43	2,19	8	1,57		

TABLA N° B11

Puntuaciones de la medida de influencia para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
28	0,49	15	0,15	26	0,06	45	-0,01	14	-0,14
47	0,48	40	0,11	24	0,05	43	-0,01	31	-0,14
20	0,39	42	0,09	25	0,04	33	-0,02	19	-0,16
9	0,30	16	0,08	30	0,03	11	-0,03	34	-0,19
27	0,29	37	0,08	4	0,03	13	-0,03	18	-0,19
21	0,28	44	0,07	39	0,02	8	-0,07	3	-0,24
35	0,27	22	0,07	36	0,01	32	-0,07	48	-0,25
6	0,24	17	0,07	1	0,00	10	-0,11	49	-0,28
29	0,18	23	0,07	7	0,00	38	-0,11	41	-0,40
12	0,17	2	0,06	46	-0,00	5	-0,13		

TABLA N° B12

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
8	0,0061	2	0,0022	35	0,0017	13	0,0010	45	0,0006
29	0,0055	9	0,0021	30	0,0016	47	0,0010	16	0,0005
31	0,0051	34	0,0021	12	0,0015	18	0,0009	32	0,0005
21	0,0034	36	0,0021	46	0,0014	39	0,0009	3	0,0004
26	0,0032	25	0,0019	11	0,0013	27	0,0008	38	0,0004
10	0,0029	1	0,0018	28	0,0013	42	0,0008	23	0,0003
15	0,0028	14	0,0018	24	0,0012	44	0,0008	48	0,0003
40	0,0028	20	0,0018	17	0,0011	5	0,0007	41	0,0002
37	0,0027	4	0,0017	22	0,0011	33	0,0007	49	0,0002
19	0,0025	6	0,0017	7	0,0010	43	0,0007		

TABLA N° B13

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de taylor

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
29	0,00	10	0,00	36	0,00	22	0,00	45	0,00
21	0,00	6	0,00	14	0,00	27	0,00	16	0,00
8	0,00	28	0,00	30	0,00	13	0,00	3	0,00
15	0,00	35	0,00	47	0,00	39	0,00	23	0,00
9	0,00	12	0,00	11	0,00	42	0,00	32	0,00
20	0,00	19	0,00	46	0,00	44	0,00	38	0,00
40	0,00	31	0,00	17	0,00	18	0,00	48	0,00
26	0,00	1	0,00	24	0,00	5	0,00	41	0,00
37	0,00	4	0,00	34	0,00	33	0,00	49	0,00
2	0,00	25	0,00	7	0,00	43	0,00		

TABLA N° B14 Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
28	0,0393	6	0,0159	14	0,0104	17	0,0058	49	0,0040
20	0,0372	10	0,0148	25	0,0096	3	0,0055	33	0,0038
47	0,0345	40	0,0145	1	0,0088	5	0,0052	43	0,0038
29	0,0312	19	0,0140	27	0,0088	13	0,0052	44	0,0038
21	0,0304	37	0,0137	4	0,0087	22	0,0052	45	0,0036
8	0,0295	35	0,0133	30	0,0082	7	0,0050	32	0,0034
31	0,0254	34	0,0120	46	0,0071	39	0,0048	38	0,0033
9	0,0250	2	0,0111	18	0,0068	48	0,0043	16	0,0025
15	0,0168	36	0,0107	11	0,0067	41	0,0040	23	0,0018
26	0,0161	12	0,0105	24	0,0059	42	0,0040		

SEGUNDO CONJUNTO DE DATOS

TABLA N° B15 Puntuaciones de la medida de influencia para la distancia de mahalanobis									
Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
90	24,51	82	6,51	33	3,72	41	2,08	42	0,44
100	21,99	20	6,37	69	3,66	48	2,07	13	0,38
64	18,59	77	6,24	60	3,63	80	2,05	63	0,35
53	18,46	88	6,18	81	3,43	18	1,95	97	0,31
74	15,46	65	5,87	24	3,35	57	1,59	95	0,30
91	14,30	84	5,86	89	3,27	23	1,58	34	0,23
98	14,11	71	5,78	86	3,24	45	1,58	68	0,23
62	12,77	11	5,28	4	3,15	1	1,53	50	0,21
75	11,13	72	5,28	59	3,04	35	1,46	40	0,19
99	10,59	79	5,12	44	2,98	47	1,39	43	0,18
94	10,56	96	5,02	8	2,80	19	1,39	37	0,17
56	8,48	92	4,89	17	2,65	5	1,23	39	0,16
21	8,47	16	4,80	31	2,56	25	1,01	66	0,14
28	7,66	32	4,54	15	2,54	27	0,94	87	0,12
78	7,17	30	4,43	6	2,46	51	0,90	49	0,05
70	7,03	73	4,34	3	2,38	12	0,85	61	0,04
85	6,83	36	4,23	29	2,36	76	0,85	55	0,04
10	6,79	26	3,94	93	2,24	14	0,83	7	0,04
83	**6,68**	22	3,86	46	2,16	58	0,70	67	0,03
54	6,65	9	3,72	38	2,15	2	0,66	52	0,01

TABLA N° B16

Puntuaciones de la medida de influencia para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
90	0,2560	84	0,0744	89	0,0303	23	-0,0099	40	-0,0693
100	0,2368	71	0,0730	86	0,0293	45	-0,0102	43	-0,0704
53	0,2084	11	0,0648	4	0,0275	1	-0,0117	37	-0,0710
74	0,1818	72	0,0648	59	0,0264	35	-0,0136	39	-0,0719
91	0,1724	79	0,0640	44	0,0245	47	-0,0143	49	-0,0848
98	0,1707	96	0,0611	8	0,0198	19	-0,0157	7	-0,0862
62	0,1583	92	0,0600	17	0,0187	5	-0,0208	55	-0,0869
75	0,1398	16	0,0574	15	0,0153	25	-0,0279	52	-0,0931
99	0,1349	32	0,0552	31	0,0149	27	-0,0309	67	-0,1115
21	0,1118	30	0,0538	6	0,0130	51	-0,0323	61	-0,1162
28	0,0999	73	0,0504	46	0,0112	12	-0,0330	87	-0,1273
78	0,0936	36	0,0484	3	0,0108	76	-0,0340	66	-0,1287
70	0,0916	22	0,0440	29	0,0102	14	-0,0343	95	-0,1410
85	0,0888	26	0,0422	93	0,0078	2	-0,0415	97	-0,1422
10	0,0873	60	0,0388	38	0,0048	42	-0,0535	58	-0,1624
83	0,0855	9	0,0386	48	0,0047	13	-0,0541	65	-0,2801
54	0,0853	33	0,0386	80	0,0045	63	-0,0586	82	-0,2891
20	0,0825	69	0,0374	41	0,0043	68	-0,0650	56	-0,3169
77	**0,0812**	81	0,0334	18	0,0035	34	-0,0662	94	-0,3421
88	0,0796	24	0,0316	57	-0,0093	50	-0,0679	64	-0,4220

TABLA N° B17

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de taylor

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
90	0,6655	79	0,1202	15	0,0341	13	0,0151	23	0,0063
100	0,3994	60	0,1182	71	0,0321	67	0,0150	40	0,0062
62	0,3405	20	0,1153	96	0,0321	57	0,0146	76	0,0060
21	0,3245	65	0,1100	54	0,0312	12	0,0145	50	0,0058
98	0,2595	85	0,0978	95	0,0289	93	0,0143	72	0,0056
91	0,2435	92	0,0975	41	0,0281	31	0,0142	39	0,0055
53	0,2262	18	0,0844	44	0,0280	68	0,0142	43	0,0053
64	0,2160	88	0,0823	47	0,0243	69	0,0125	5	0,0051
99	0,2008	75	0,0758	16	0,0216	4	0,0123	34	0,0050
46	0,1786	84	0,0672	97	0,0215	14	0,0119	51	0,0050
77	0,1632	73	0,0664	6	0,0200	3	0,0101	27	0,0038
82	0,1589	17	0,0643	89	0,0197	29	0,0101	63	0,0037
94	0,1572	74	0,0635	86	0,0191	52	0,0098	19	0,0035
30	0,1540	80	0,0629	66	0,0175	25	0,0090	1	0,0033
78	0,1498	59	0,0622	87	0,0173	55	0,0086	8	0,0032
22	0,1466	48	0,0457	83	0,0172	11	0,0080	35	0,0031
32	0,1328	36	0,0436	9	0,0167	2	0,0078	42	0,0031
70	0,1322	81	0,0369	33	0,0167	7	0,0076	45	0,0026
56	**0,1225**	10	0,0354	61	0,0166	49	0,0072	26	0,0018
28	0,1220	58	0,0353	24	0,0153	37	0,0065	38	0,0002

TABLA N° B18

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
90	0,4053	20	0,0676	58	0,0200	13	0,0091	37	0,0042
100	0,2530	70	0,0672	81	0,0193	61	0,0091	40	0,0040
62	0,1962	85	0,0621	44	0,0169	12	0,0088	23	0,0039
21	0,1784	79	0,0607	41	0,0168	67	0,0088	72	0,0039
98	0,1653	60	0,0569	54	0,0168	31	0,0087	50	0,0037
91	0,1573	88	0,0559	96	0,0161	86	0,0082	39	0,0035
53	0,1366	75	0,0534	89	0,0151	57	0,0077	43	0,0034
64	0,1271	92	0,0475	95	0,0149	4	0,0076	5	0,0032
99	0,1207	18	0,0457	47	0,0144	14	0,0074	34	0,0032
82	0,0955	74	0,0414	97	0,0141	68	0,0064	51	0,0032
94	0,0913	17	0,0369	16	0,0133	29	0,0063	63	0,0031
77	0,0884	84	0,0346	93	0,0123	3	0,0062	27	0,0024
46	0,0860	73	0,0331	6	0,0122	52	0,0057	19	0,0022
30	0,0852	59	0,0299	83	0,0116	25	0,0056	1	0,0021
22	0,0802	80	0,0299	66	0,0115	55	0,0051	8	0,0020
78	0,0774	48	0,0265	87	0,0106	11	0,0050	42	0,0020
32	0,0748	36	0,0262	69	0,0104	2	0,0049	35	0,0019
56	0,0732	71	0,0248	9	0,0103	7	0,0049	45	0,0016
28	**0,0722**	10	0,0218	33	0,0103	49	0,0046	26	0,0011
65	0,0703	15	0,0203	24	0,0094	76	0,0045	38	0,0002

TABLA N° B19

Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
64	0,0996	77	0,0089	30	0,0060	13	0,0021	7	0,0015
90	0,0585	79	0,0085	88	0,0060	15	0,0021	10	0,0015
100	0,0349	93	0,0085	22	0,0059	41	0,0021	49	0,0015
94	0,0252	67	0,0082	85	0,0059	12	0,0020	1	0,0014
53	0,0168	61	0,0081	84	0,0058	14	0,0019	19	0,0014
21	0,0139	92	0,0080	58	0,0056	36	0,0019	23	0,0014
56	0,0139	99	0,0078	96	0,0055	2	0,0018	35	0,0014
62	0,0138	46	0,0077	28	0,0052	25	0,0018	29	0,0013
98	0,0115	81	0,0077	32	0,0052	37	0,0017	31	0,0013
91	0,0110	73	0,0076	74	0,0048	40	0,0017	45	0,0013
80	0,0105	78	0,0075	71	0,0046	44	0,0017	3	0,0012
60	0,0104	87	0,0075	20	0,0045	50	0,0017	4	0,0010
68	0,0101	66	0,0074	18	0,0043	6	0,0016	9	0,0010
63	0,0097	86	0,0073	72	0,0042	27	0,0016	24	0,0010
76	0,0097	89	0,0073	54	0,0036	34	0,0016	33	0,0010
82	0,0097	70	0,0070	17	0,0032	39	0,0016	38	0,0010
55	0,0093	95	0,0069	83	0,0031	42	0,0016	16	0,0009
57	0,0093	69	0,0066	75	0,0029	43	0,0016	8	0,0008
59	0,0092	97	0,0066	48	0,0027	51	0,0016	11	0,0003
52	0,0091	65	0,0063	47	0,0022	5	0,0015	26	0,0003

TERCER CONJUNTO DE DATOS

TABLA N° B20

Puntuaciones de la medida de influencia para la distancia de Mahalanobis

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
34	65,11	30	15,39	10	5,83	63	2,75	13	0,40
84	46,76	88	13,50	3	5,60	32	2,71	79	0,40
21	43,84	76	13,08	36	5,53	18	2,60	22	0,39
69	39,79	49	13,02	65	5,30	75	2,57	59	0,34
23	38,96	67	12,47	62	4,87	83	2,27	16	0,22
89	32,76	29	11,88	54	4,58	55	2,10	99	0,22
19	32,55	57	11,15	8	4,44	60	2,04	53	0,19
80	30,96	14	10,93	40	4,27	91	1,74	20	0,18
28	30,34	98	10,62	45	4,05	94	1,63	39	0,11
78	30,06	5	9,47	44	3,86	2	1,58	46	0,09
77	29,57	7	8,86	24	3,80	58	1,45	25	0,09
74	23,06	4	8,15	73	3,63	81	1,37	68	0,08
35	20,81	6	8,09	95	3,61	87	1,30	31	0,08
61	20,61	38	7,93	37	3,55	96	1,28	64	0,07
51	19,22	90	7,82	12	3,33	47	0,94	48	0,04
70	17,72	41	7,66	56	3,07	50	0,76	11	0,04
17	16,47	97	6,99	15	3,01	33	0,71	26	0,03
82	16,16	92	6,37	72	2,84	86	0,62	1	0,02
100	15,95	27	5,92	52	2,83	9	0,46	43	0,00
85	15,66	42	5,85	93	2,83	66	0,46	71	0,00

TABLA N° B21

Puntuaciones de la medida de influencia para la probabilidad de mala clasificación

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
34	0,27	100	0,05	48	0,02	94	0,02	15	0,01
84	0,18	49	0,04	31	0,02	7	0,02	12	0,01
69	0,18	88	0,04	39	0,02	91	0,02	95	0,01
21	0,18	76	0,04	25	0,02	2	0,02	37	0,01
23	0,15	67	0,03	46	0,02	55	0,02	73	0,01
19	0,13	29	0,03	86	0,02	60	0,02	97	0,01
89	0,12	57	0,03	20	0,02	83	0,01	24	0,01
80	0,12	14	0,03	16	0,02	75	0,01	92	0,01
28	0,11	71	0,03	98	0,02	4	0,01	45	0,01
78	0,11	64	0,03	9	0,02	38	0,01	54	0,01
77	0,11	68	0,03	22	0,02	63	0,01	44	0,00
74	0,08	53	0,03	13	0,02	6	0,01	40	0,00
35	0,07	99	0,03	50	0,02	52	0,01	62	0,00
51	0,07	59	0,02	96	0,02	93	0,01	27	0,00
61	0,07	66	0,02	5	0,02	72	0,01	10	0,00
70	0,06	79	0,02	33	0,02	41	0,01	42	0,00
82	0,06	43	0,02	81	0,02	18	0,01	36	0,00
85	0,05	1	0,02	47	0,02	56	0,01	3	0,00
17	0,05	11	0,02	58	0,02	32	0,01	65	0,00
30	0,05	26	0,02	87	0,02	90	0,01	8	0,00

TABLA N° B22

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación

Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones
69	0,42	38	0,04	67	0,02	96	0,01	46	0,00
19	0,27	73	0,04	27	0,02	37	0,01	25	0,00
82	0,24	100	0,04	60	0,02	94	0,01	31	0,00
85	0,24	61	0,04	3	0,02	42	0,01	1	0,00
80	0,18	77	0,04	15	0,02	13	0,01	63	0,00
34	0,17	30	0,04	5	0,02	81	0,01	59	0,00
21	0,15	74	0,04	10	0,01	52	0,01	71	0,00
35	0,12	88	0,03	54	0,01	93	0,01	68	0,00
70	0,12	78	0,03	14	0,01	29	0,01	43	0,00
57	0,11	36	0,03	87	0,01	40	0,01	66	0,00
51	0,10	41	0,03	18	0,01	33	0,00	99	0,00
84	0,10	97	0,03	91	0,01	22	0,00	11	0,00
23	0,10	7	0,03	32	0,01	16	0,00	26	0,00
49	0,08	4	0,02	24	0,01	62	0,00	48	0,00
92	0,07	56	0,02	83	0,01	79	0,00	64	0,00
76	0,06	8	0,02	98	0,01	12	0,00	39	0,00
17	0,06	95	0,02	86	0,01	20	0,00	9	0,00
65	0,05	44	0,02	72	0,01	75	0,00	47	0,00
28	0,05	6	0,02	58	0,01	2	0,00	53	0,00
89	0,04	90	0,02	55	0,01	45	0,00	50	0,00

TABLA N° B23

Puntuaciones de la medida de influencia alternativa para la probabilidad de mala clasificación mediante aproximación de taylor

Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones	Obs.	Puntuaciones
69	0,15	61	0,02	67	0,01	37	0,00	46	0,00
19	0,10	100	0,02	27	0,01	94	0,00	25	0,00
82	0,08	38	0,02	3	0,01	96	0,00	31	0,00
85	0,08	77	0,02	5	0,01	42	0,00	1	0,00
34	0,07	30	0,02	15	0,01	13	0,00	63	0,00
80	0,07	74	0,02	60	0,01	29	0,00	59	0,00
21	0,06	73	0,02	10	0,01	52	0,00	68	0,00
35	0,05	88	0,01	14	0,01	93	0,00	71	0,00
70	0,05	78	0,01	54	0,01	81	0,00	43	0,00
84	0,04	41	0,01	18	0,01	33	0,00	66	0,00
57	0,04	36	0,01	32	0,00	40	0,00	99	0,00
51	0,04	97	0,01	87	0,00	22	0,00	11	0,00
23	0,04	7	0,01	91	0,00	16	0,00	26	0,00
49	0,03	4	0,01	24	0,00	79	0,00	64	0,00
92	0,03	8	0,01	83	0,00	62	0,00	39	0,00
76	0,02	56	0,01	98	0,00	12	0,00	48	0,00
17	0,02	95	0,01	72	0,00	2	0,00	9	0,00
28	0,02	44	0,01	55	0,00	20	0,00	53	0,00
65	0,02	6	0,01	86	0,00	75	0,00	47	0,00
89	0,02	90	0,01	58	0,00	45	0,00	50	0,00

TABLA N° B24

Puntuaciones de la medida de influencia para las puntuaciones de la función discriminante lineal de Fisher

Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones	Obs.	Puntu aciones
69	0,31	76	0,01	7	0,00	48	0,00	16	0,00
34	0,27	38	0,01	96	0,00	39	0,00	13	0,00
19	0,11	74	0,01	2	0,00	47	0,00	79	0,00
21	0,11	61	0,01	52	0,00	63	0,00	44	0,00
84	0,11	100	0,01	93	0,00	71	0,00	56	0,00
23	0,08	17	0,01	10	0,00	43	0,00	60	0,00
80	0,07	65	0,01	59	0,00	41	0,00	87	0,00
85	0,06	88	0,01	12	0,00	46	0,00	22	0,00
82	0,06	4	0,01	66	0,00	8	0,00	91	0,00
51	0,05	36	0,01	68	0,00	90	0,00	6	0,00
28	0,04	58	0,01	75	0,00	1	0,00	42	0,00
89	0,03	27	0,01	99	0,00	86	0,00	33	0,00
70	0,03	5	0,01	9	0,00	31	0,00	29	0,00
35	0,03	3	0,01	53	0,00	18	0,00	98	0,00
57	0,02	54	0,01	64	0,00	25	0,00	32	0,00
78	0,02	73	0,01	40	0,00	62	0,00	94	0,00
49	0,02	81	0,01	50	0,00	20	0,00	83	0,00
30	0,02	97	0,01	67	0,00	24	0,00	14	-
77	0,02	37	0,01	11	0,00	45	0,00	15	-
92	0,01	72	0,01	26	0,00	95	0,00	55	-

D. PROGRAMAS PARA CALCULAR LAS MEDIDAS

SIMULACIÓN, PARA GENERAR DOS POBLACIONES NORMALES MEDIANTE EL MÉTODO DE MONTECARLO

```
# Elegir método
# Método 1 = "1"
# Método 2 = "2"
# ------------------------------------------------------------------------------
library(clusterGeneration) # debe estar descargada en la PC.
library(MASS)
# ------------------------------------------------------------------------------
METODO = 1
n1 = 50
mu1 = c(1,2)
n2 = 50
mu2 = c(2,3)
p = length(mu2)

if (METODO==1){
  sigma = genPositiveDefMat("unifcorrmat",dim=p)$Sigma
  poblacion1 = mvrnorm(n1,mu1,sigma)
  poblacion2 = mvrnorm(n2,mu2,sigma)
}
else{
  P1 = matrix(0,nrow=n1,ncol=p)
  P2 = matrix(0,nrow=n2,ncol=p)
  for (i in 1:p){
          P1[,i] = rnorm(n1,0,1)
          P2[,i] = rnorm(n2,0,1)
  }
  sigma = genPositiveDefMat("unifcorrmat",dim=p)$Sigma
  Lt = chol(sigma)
```

```
  poblacion1=matrix(rep(mu1,n1),n1)+(P1%*%Lt)
  poblacion2=matrix(rep(mu2,n2),n2)+(P2%*%Lt)
}
```

PRUEBA M-BOX DE IGUALDAD DE MATRICES DE COVARIANZAS

```
sigmaest1 = cov(poblacion1)
sigmaest2 = cov(poblacion2)
sigmaest = ((n1-1)*sigmaest1+(n2-1)*sigmaest2)/(n1+n2-2)
muest1 = colMeans(poblacion1)
muest2 = colMeans(poblacion2)
p = length(muest2)
n = n1+n2
M=(n-2)*log(det(sigmaest))-((n1-1)*log(det(sigmaest1))+(n2-1)*log(det(sigmaest2)))
gl = p*(p+1)/2
CHI = qchisq(0.95,gl,lower.tail = TRUE, log.p = FALSE)
```

PRUEBA DE MARDIA DE MULTINORMALIDAD

CÁLCULO DE LAS DISTANCIAS DE MAHALANOBIS

```
d1 = matrix(0,ncol=n1,nrow=n1)
d2 = matrix(0,ncol=n2,nrow=n2)
for ( i in 1:n1){
    for (j in 1:n1){
   d1[i,j] = (poblacion1[i,]-muest1)%*%solve(sigma)%*%(poblacion1[j,]-muest1)
          }
}
for ( i in 1:n2){
          for (j in 1:n2){
          d2[i,j] = (poblacion2[i,]-muest2)%*%solve(sigma)%*%(poblacion2[j,]-muest2)
          }
}
```

PRUEBA DE SIMETRIA

```
AP1 = sum(d1^3)/(n1^2)
AP2 = sum(d2^3)/(n2^2)
```

```
f = p*(p+1)*(p+2)/6
CHI = qchisq(0.975,f,lower.tail = TRUE, log.p = FALSE)
M1 = n1*AP1/6
M2 = n2*AP2/6
```

PRUEBA PARA LA CURTOSIS

```
kp1 = sum(diag(d1^2))/n1
Z1 = qnorm(0.975, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
Ekp1 = (kp1-(p*(p+2)))/sqrt((8*p*(p+2))/n1)

kp2 = sum(diag(d2^2))/n2
Z2 = qnorm(0.975, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
Ekp2 = (kp2-(p*(p+2)))/sqrt((8*p*(p+2))/n2)
```

Regla de Decisión

```
if (M<CHI) print("Matrices de Covarianza Iguales")
if (M1<CHI){
        if(abs(Ekp1)<Z1){
                print("La población 1 es Normal Multivariada")
        }
}
if(M2<CHI){
        if (abs(Ekp2)<Z2){
                print("La población 2 es Normal Multivariada")
        }
}
```

EXPORTACIÓN DE DATOS A EXCEL

```
write.table(poblacion1,"C:/Poblacion1.xls",sep="\t",col.names=TRUE,  row.names=FALSE,
quote=TRUE, na="NA")
write.table(poblacion2,"C:/Poblacion2.xls",sep="\t",col.names=TRUE,  row.names=FALSE,
quote=TRUE, na="NA")
```

MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL PARA DOS GRUPOS.

Primer conjunto de datos

```
% PRIMERA MEDIDA: FUNCION DE INFLUENCIA DE LA DISTANCIA DE MAHALANOBIS
clear;
clc;
X1=xlsread('Sobrevivieron.xls');       % LECTURA POBLACION 1
[n1,v1]=size(X1);                  % N° INDIVIDUOS Y N° VARIABLES POBLACION 1
X2=xlsread('Murieron.xls');       % LECTURA POBLACION 2
[n2,v2]=size(X2);                  % N° INDIVIDUOS Y N° VARIABLES POBLACION 2
mX1=mean(X1);                      % MEDIA POBLACION 1
mX2=mean(X2);                      % MEDIA POBLACION 2
sX1=cov(X1);                      % COVARIANZA POBLACION 1
sX2=cov(X2);                      % COVARIANZA POBLACION 2
w1=(n1-1)/(n1+n2-2);               % PESO 1
w2=(n2-1)/(n1+n2-2);               % PESO 2
sU=w1*sX1+w2*sX2;                  % COVARIANZA CONJUNTA
a=sU\(mX1-mX2)';                   % ALFA
for i=1:n1
   trin1(i)=a'*(X1(i,:)-mX1)';     % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 1
end
for i=1:n2
   trin2(i)=a'*(mX2-X2(i,:))';     % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 2
% En esta sentencia difiere de tu programa. Las puntuaciones difieren en algunos puntos de la grafica.
end
```

```
trin=[trin1 trin2]';              % DISTANCIA CONJUNTA
for i=1:n1+n2
   IM(i)=w1*(trin(i)-inv(w1)).^2;  % PRIMERA MEDIDA
end
IM=IM';
IM
%plot(IM);

% SEGUNDA MEDIDA: MEDIDA BASADA EN LA PROBABILIDAD DE MALA CLASIFICACION
CON OMISION

D2=((mX1-mX2)/sU)*(mX1-mX2)';
for i=1:n1
   if i==1
      X1i=X1(2:n1,:);
   else
      X1i1=X1(1:i-1,:);
      X1i2=X1(i+1:n1,:);
      X1i=[X1i1;X1i2];
   end
   mX1i(i,:)=mean(X1i);
   sX1i(:,:,i)=cov(X1i);
   D21i(i)=(mX1i(i,:)-mX2)*(inv(sU))*(mX1i(i,:)-mX2)';
end
IMP1=(n1-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D21i)));
for i=1:n2
   if i==1
```

```
    X2i=X2(2:n2,:);
  else
    X2i1=X2(1:i-1,:);
    X2i2=X2(i+1:n2,:);
    X2i=[X2i1;X2i2];
  end
  mX2i(i,:)=mean(X2i);
  sX2i(:,:,i)=cov(X2i);
  D22i(i)=(mX2i(i,:)-mX1)*(inv(sU))*(mX2i(i,:)-mX1)';
end
IMP2=(n2-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D22i)));
IMP=[IMP1';IMP2'];
IMP
%plot(IMP);
% TERCERA MEDIDA: % MEDIDA BASADA EN LA PROBABILIDAD DE MALA CLASIFICACION CON OMISION

w1i1=((n1-1)-1)/((n1-1)+n2-2);
w2i1=(n2-1)/((n1-1)+n2-2);
for i=1:n1
  sU1i=w1i1*sX1i(:,:,i)+w2i1*sX2;
  ai1(:,:,i)=inv(sU1i)*(mX1i(i,:)-mX2)';
  G21=ai1(:,:,i)'*sU*ai1(:,:,i);
  P11=normcdf((-ai1(:,:,i)'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
  P21=normcdf((-ai1(:,:,i)'*(mX1-mX2)'+ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
  DMP1(i)=0.5*(P11+P21)-normcdf(-0.5*sqrt(D2));
end
```

```
w1i2=(n1-1)/(n1+(n2-1)-2);
w2i2=((n2-1)-1)/(n1+(n2-1)-2);
for i=1:n2
   sU2i=w1i2*sX1+w2i1*sX2i(:,:,i);
   ai2(:,:,i)=inv(sU2i)*(mX1-mX2i(i,:))';
   G22=ai2(:,:,i)'*sU*ai2(:,:,i);
   P12=normcdf((-ai2(:,:,i)'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
   P22=normcdf((-ai2(:,:,i)'*(mX1-mX2)'+ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
   DMP2(i)=0.5*(P12+P22)-normcdf(-0.5*sqrt(D2));
end
DMP=[DMP1';DMP2'];
%plot(DMP);
DMP
% CUARTA MEDIDA:

k1=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n1-1)^2);
for i=1:n1
   d21=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';
   DMP11(i)=k1*((1-w1*trin1(i))^2*(d21-trin1(i)^2/D2)+0.25*trin1(i)^2);
end
k2=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n2-1)^2);
for i=1:n2
   d22=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
   DMP12(i)=k2*((1-w2*trin2(i))^2*(d22-trin2(i)^2/D2)+0.25*trin2(i)^2);
end
DMPi=[DMP11';DMP12'];
%plot(DMPi);
```

```
DMPi
% QUINTA MEDIDA:


t1=(n1-1)/((n1-1)+n2);
for i=1:n1
   V1=(a-ai1(:,:,i))'*sU*(a-ai1(:,:,i));
   B11=((a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
   B21=(-1*(a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
   E21(i)=t1*B11^2+(1-t1)*B21^2+V1;
end
t2=(n2-1)/(n1+(n2-1));
for i=1:n2
   V2=(a-ai2(:,:,i))'*sU*(a-ai2(:,:,i));
   B12=((a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   B22=(-1*(a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   E22(i)=t2*B12^2+(1-t2)*B22^2+V2;
end
E2=[E21';E22'];
E2
%plot(E2);
%MEDIDAS ADICIONALES
%PRIMERA MEDIDA ADICIONAL
for i=1:n1
   IM1(i)=w1*(trin1(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL PRIMER GRUPO
end
IM1=IM1';
%plot(IM1);
```

```
for i=1:n2
   IM2(i)=w1*(trin2(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL SEGUNDO GRUPO
end
IM2=IM2';
%plot(IM2);
%SEGUNDA MEDIDA ADICIONAL
for i=1:n1
   d21(i)=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';% SEGUNDA MEDIDA ADICIONAL PRIMER GRUPO
end
d21i=d21';
%plot(d21);
for i=1:n2
   d22(i)=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
end
d22i=d22';
%plot(d22);

   % MOSTRANDO LAS GRAFICAS

subplot(3,3,1);
plot(IM);
title('PRIMERA MEDIDA');
subplot(3,3,2);
plot(IMP);
title('SEGUNDA MEDIDA');
subplot(3,3,3);
```

```
plot(DMP);
title('TERCERA MEDIDA');
subplot(3,3,4);
plot(DMPi);
title('CUARTA MEDIDA');
subplot(3,3,5);
plot(E2);
title('QUINTA MEDIDA');
subplot(3,3,6);
plot(IM1);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,7);
plot(IM2);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 2');
subplot(3,3,8);
plot(d21i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,9);
plot(d22i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 2');
```

Segundo conjunto de datos

```
% PRIMERA MEDIDA: FUNCION DE INFLUENCIA DE LA DISTANCIA DE MAHALANOBIS

clear;
clc;
X1=xlsread('Abundante.xls');       % LECTURA POBLACION 1
```

```
[n1,v1]=size(X1);              % N° INDIVIDUOS Y N° VARIABLES POBLACION 1
X2=xlsread('Escasa.xls');      % LECTURA POBLACION 2
[n2,v2]=size(X2);              % N° INDIVIDUOS Y N° VARIABLES POBLACION 2
mX1=mean(X1);                  % MEDIA POBLACION 1
mX2=mean(X2);                  % MEDIA POBLACION 2
sX1=cov(X1);                   % COVARIANZA POBLACION 1
sX2=cov(X2);                   % COVARIANZA POBLACION 2
w1=(n1-1)/(n1+n2-2);           % PESO 1
w2=(n2-1)/(n1+n2-2);           % PESO 2
sU=w1*sX1+w2*sX2;              % COVARIANZA CONJUNTA
a=sU\(mX1-mX2)';               % ALFA
for i=1:n1
   trin1(i)=a'*(X1(i,:)-mX1)';    % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 1
end
for i=1:n2
   trin2(i)=a'*(mX2-X2(i,:))';    % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 2
% En esta sentencia difiere de tu programa. Las puntuaciones difieren en algunos puntos de la grafica.
end
trin=[trin1 trin2]';           % DISTANCIA CONJUNTA
for i=1:n1+n2
   IM(i)=w1*(trin(i)-inv(w1)).^2;  % PRIMERA MEDIDA
end
IM=IM';
IM
%plot(IM);
```

```
% SEGUNDA MEDIDA: MEDIDA BASADA EN LA PROBABILIDAD DE MALA CLASIFICACION
CON OMISION


D2=((mX1-mX2)/sU)*(mX1-mX2)';
for i=1:n1
   if i==1
      X1i=X1(2:n1,:);
   else
      X1i1=X1(1:i-1,:);
      X1i2=X1(i+1:n1,:);
      X1i=[X1i1;X1i2];
   end
   mX1i(i,:)=mean(X1i);
   sX1i(:,:,i)=cov(X1i);
   D21i(i)=(mX1i(i,:)-mX2)*(inv(sU))*(mX1i(i,:)-mX2)';
end
IMP1=(n1-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D21i)));
for i=1:n2
   if i==1
      X2i=X2(2:n2,:);
   else
      X2i1=X2(1:i-1,:);
      X2i2=X2(i+1:n2,:);
      X2i=[X2i1;X2i2];
   end
   mX2i(i,:)=mean(X2i);
   sX2i(:,:,i)=cov(X2i);
```

```
   D22i(i)=(mX2i(i,:)-mX1)*(inv(sU))*(mX2i(i,:)-mX1)';
end
IMP2=(n2-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D22i)));
IMP=[IMP1';IMP2'];
%plot(IMP);
IMP
% TERCERA MEDIDA: % MEDIDA BASADA EN LA PROBABILIDAD DE MALA CLASIFICACION CON OMISION

w1i1=((n1-1)-1)/((n1-1)+n2-2);
w2i1=(n2-1)/((n1-1)+n2-2);
for i=1:n1
   sU1i=w1i1*sX1i(:,:,i)+w2i1*sX2;
   ai1(:,:,i)=inv(sU1i)*(mX1i(i,:)-mX2)';
   G21=ai1(:,:,i)'*sU*ai1(:,:,i);
   P11=normcdf((-ai1(:,:,i)'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
   P21=normcdf((-ai1(:,:,i)'*(mX1-mX2)'+ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
   DMP1(i)=0.5*(P11+P21)-normcdf(-0.5*sqrt(D2));
end
w1i2=(n1-1)/(n1+(n2-1)-2);
w2i2=((n2-1)-1)/(n1+(n2-1)-2);
for i=1:n2
   sU2i=w1i2*sX1+w2i1*sX2i(:,:,i);
   ai2(:,:,i)=sU2i\(mX1-mX2i(i,:))';
   G22=ai2(:,:,i)'*sU*ai2(:,:,i);
   P12=normcdf((-ai2(:,:,i)'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
   P22=normcdf((-ai2(:,:,i)'*(mX1-mX2)'+ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
```

```
   DMP2(i)=0.5*(P12+P22)-normcdf(-0.5*sqrt(D2));
end
DMP=[DMP1';DMP2'];
%plot(DMP);
DMP
% CUARTA MEDIDA:

k1=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n1-1)^2);
for i=1:n1
   d21=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';
   DMP11(i)=k1*((1-w1*trin1(i))^2*(d21-trin1(i)^2/D2)+0.25*trin1(i)^2);
end
k2=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n2-1)^2);
for i=1:n2
   d22=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
   DMP12(i)=k2*((1-w2*trin2(i))^2*(d22-trin2(i)^2/D2)+0.25*trin2(i)^2);
end
DMPi=[DMP11';DMP12'];
%plot(DMPi);
DMPi
% QUINTA MEDIDA:

t1=(n1-1)/((n1-1)+n2);
for i=1:n1
   V1=(a-ai1(:,:,i))'*sU*(a-ai1(:,:,i));
   B11=((a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
   B21=(-1*(a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
```

```
   E21(i)=t1*B11^2+(1-t1)*B21^2+V1;
end
t2=(n2-1)/(n1+(n2-1));
for i=1:n2
   V2=(a-ai2(:,:,i))'*sU*(a-ai2(:,:,i));
   B12=((a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   B22=(-1*(a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   E22(i)=t2*B12^2+(1-t2)*B22^2+V2;
end
E2=[E21';E22']
%plot(E2);
E2
%MEDIDAS ADICIONALES
%PRIMERA MEDIDA ADICIONAL
for i=1:n1
   IM1(i)=w1*(trin1(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL PRIMER GRUPO
end
IM1=IM1';
%plot(IM1);
for i=1:n2
   IM2(i)=w1*(trin2(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL SEGUNDO GRUPO
end
IM2=IM2';
%plot(IM2);
%SEGUNDA MEDIDA ADICIONAL
for i=1:n1
```

```
   d21(i)=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';% SEGUNDA MEDIDA ADICIONAL PRIMER GRUPO
end
d21i=d21';
%plot(d21);
for i=1:n2
   d22(i)=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
end
d22i=d22';
%plot(d22);

   % MOSTRANDO LAS GRAFICAS

subplot(3,3,1);
plot(IM);
title('PRIMERA MEDIDA');
subplot(3,3,2);
plot(IMP);
title('SEGUNDA MEDIDA');
subplot(3,3,3);
plot(DMP);
title('TERCERA MEDIDA');
subplot(3,3,4);
plot(DMPi);
title('CUARTA MEDIDA');
subplot(3,3,5);
plot(E2);
```

```
title('QUINTA MEDIDA');
subplot(3,3,6);
plot(IM1);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,7);
plot(IM2);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 2');
subplot(3,3,8);
plot(d21i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,9);
plot(d22i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 2');
```

Tercer conjunto de datos

```
%MEDIDAS DE INFLUENCIA EN EL ANÁLISIS DISCRIMINANTE LINEAL PARA DOS GRUPOS
clear;
clc;
X1=xlsread('Versicolor.xls');      % LECTURA POBLACION 1
[n1,v1]=size(X1);                  % N° INDIVIDUOS Y N° VARIABLES POBLACION 1
X2=xlsread('Virginica.xls');       % LECTURA POBLACION 2
[n2,v2]=size(X2);                  % N° INDIVIDUOS Y N° VARIABLES POBLACION 2
mX1=mean(X1);                      % MEDIA POBLACION 1
mX2=mean(X2);                      % MEDIA POBLACION 2
sX1=cov(X1);                       % COVARIANZA POBLACION 1
sX2=cov(X2);                       % COVARIANZA POBLACION 2
```

```
w1=(n1-1)/(n1+n2-2);              % PESO 1
w2=(n2-1)/(n1+n2-2);              % PESO 2
sU=w1*sX1+w2*sX2;                 % COVARIANZA CONJUNTA
a=sU\(mX1-mX2)';                  % ALFA

% MEDIDA DE INFLUENCIA PARA DISTANCIA DE MAHALANOBIS
for i=1:n1
   trin1(i)=a'*(X1(i,:)-mX1)';    % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 1
end
for i=1:n2
   trin2(i)=a'*(mX2-X2(i,:))';    % DISTANCIA CON RESPECTO A LA MEDIA POBLACION 2
% En esta sentencia difiere de tu programa. Las puntuaciones difieren en algunos puntos de la grafica.
end
trin=[trin1 trin2]';              % DISTANCIA CONJUNTA
for i=1:n1+n2
   IM(i)=w1*(trin(i)-inv(w1)).^2;  % PRIMERA MEDIDA
end
IM=IM';
%plot(IM);

% MEDIDA PARA PROBABILIDAD DE MALA CLASIFICACION CON OMISION

D2=((mX1-mX2)/sU)*(mX1-mX2)';
for i=1:n1
   if i==1
      X1i=X1(2:n1,:);
```

```
    else
        X1i1=X1(1:i-1,:);
        X1i2=X1(i+1:n1,:);
        X1i=[X1i1;X1i2];
    end
    mX1i(i,:)=mean(X1i);
    sX1i(:,:,i)=cov(X1i);
    D21i(i)=(mX1i(i,:)-mX2)*(inv(sU))*(mX1i(i,:)-mX2)';
end
IMP1=(n1-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D21i)));
for i=1:n2
    if i==1
        X2i=X2(2:n2,:);
    else
        X2i1=X2(1:i-1,:);
        X2i2=X2(i+1:n2,:);
        X2i=[X2i1;X2i2];
    end
    mX2i(i,:)=mean(X2i);
    sX2i(:,:,i)=cov(X2i);
    D22i(i)=(mX2i(i,:)-mX1)*(inv(sU))*(mX2i(i,:)-mX1)';
end
IMP2=(n2-1)*(normcdf(-0.5*sqrt(D2))-normcdf(-0.5*sqrt(D22i)));
IMP=[IMP1';IMP2'];
%plot(IMP);

% MEDIDA  ALTERNATIVA PARA PROBABILIDAD DE MALA CLASIFICACION CON OMISION
```

```
w1i1=((n1-1)-1)/((n1-1)+n2-2);
w2i1=(n2-1)/((n1-1)+n2-2);
for i=1:n1
   sU1i=w1i1*sX1i(:,:,i)+w2i1*sX2;
   ai1(:,:,i)=inv(sU1i)*(mX1i(i,:)-mX2)';
   G21=ai1(:,:,i)'*sU*ai1(:,:,i);
   P11=normcdf((-ai1(:,:,i)'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
   P21=normcdf((-ai1(:,:,i)'*(mX1-mX2)'+ai1(:,:,i)'*(mX1-mX1i(i,:))')/(2*sqrt(G21)));
   DMP1(i)=0.5*(P11+P21)-normcdf(-0.5*sqrt(D2));
end
w1i2=(n1-1)/(n1+(n2-1)-2);
w2i2=((n2-1)-1)/(n1+(n2-1)-2);
for i=1:n2
   sU2i=w1i2*sX1+w2i1*sX2i(:,:,i);
   ai2(:,:,i)=sU2i\(mX1-mX2i(i,:))';
   G22=ai2(:,:,i)'*sU*ai2(:,:,i);
   P12=normcdf((-ai2(:,:,i)'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
   P22=normcdf((-ai2(:,:,i)'*(mX1-mX2)'+ai2(:,:,i)'*(mX2-mX2i(i,:))')/(2*sqrt(G22)));
   DMP2(i)=0.5*(P12+P22)-normcdf(-0.5*sqrt(D2));
end
DMP=[DMP1';DMP2'];
%plot(DMP);

%MEDIDA  ALTERNATIVA PARA PROBABILIDAD DE MALA CLASIFICACION CON OMISION
%MEDIANTE APROXIMACION DE TAYLOR
```

```
k1=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n1-1)^2);
for i=1:n1
   d21=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';
   DMP11(i)=k1*((1-w1*trin1(i))^2*(d21-trin1(i)^2/D2)+0.25*trin1(i)^2);
end
k2=normcdf(-0.5*sqrt(D2))/(4*sqrt(D2)*(n2-1)^2);
for i=1:n2
   d22=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
   DMP12(i)=k2*((1-w2*trin2(i))^2*(d22-trin2(i)^2/D2)+0.25*trin2(i)^2);
end
DMPi=[DMP11';DMP12'];
%plot(DMPi);

% MEDIDA DE INFLUENCIA PARA LAS PUNTUACIONES DE LA LA FUNCIÒN DISCRIMINANTE
% LINEAL DE FISHER

t1=(n1-1)/((n1-1)+n2);
for i=1:n1
   V1=(a-ai1(:,:,i))'*sU*(a-ai1(:,:,i));
   B11=((a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
   B21=(-1*(a-ai1(:,:,i))'*(mX1-mX2)'-ai1(:,:,i)'*(mX1-mX1i(i,:))')*0.5;
   E21(i)=t1*B11^2+(1-t1)*B21^2+V1;
end
t2=(n2-1)/(n1+(n2-1));
for i=1:n2
   V2=(a-ai2(:,:,i))'*sU*(a-ai2(:,:,i));
```

```
   B12=((a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   B22=(-1*(a-ai2(:,:,i))'*(mX1-mX2)'-ai2(:,:,i)'*(mX2-mX2i(i,:))')*0.5;
   E22(i)=t2*B12^2+(1-t2)*B22^2+V2;
end
E2=[E21';E22'];
%plot(E2);
%MEDIDAS ADICIONALES
%PRIMERA MEDIDA ADICIONAL
for i=1:n1
   IM1(i)=w1*(trin1(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL PRIMER GRUPO
end
IM1=IM1'
%plot(IM1);
for i=1:n2
   IM2(i)=w1*(trin2(i)-inv(w1)).^2;  % PRIMERA MEDIDA ADICIONAL SEGUNDO GRUPO
end
IM2=IM2'
%plot(IM2);
%SEGUNDA MEDIDA ADICIONAL
for i=1:n1
   d21(i)=(X1(i,:)-mX1)*inv(sU)*(X1(i,:)-mX1)';% SEGUNDA MEDIDA ADICIONAL PRIMER
GRUPO
end
d21i=d21'
%plot(d21);
for i=1:n2
   d22(i)=(X2(i,:)-mX2)*inv(sU)*(X2(i,:)-mX2)';
```

```
end
d22i=d22'
%plot(d22);

   % MOSTRANDO LAS GRAFICAS

subplot(3,3,1);
plot(IM);
title('PRIMERA MEDIDA');
subplot(3,3,2);
plot(IMP);
title('SEGUNDA MEDIDA');
subplot(3,3,3);
plot(DMP);
title('TERCERA MEDIDA');
subplot(3,3,4);
plot(DMPi);
title('CUARTA MEDIDA');
subplot(3,3,5);
plot(E2);
title('QUINTA MEDIDA');
subplot(3,3,6);
plot(IM1);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,7);
plot(IM2);
title('PRIMERA MEDIDA ADICIONAL:GRUPO 2');
```

```
subplot(3,3,8);
plot(d21i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 1');
subplot(3,3,9);
plot(d22i);
title('SEGUNDA MEDIDA ADICIONAL:GRUPO 2');
```